1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Human 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

生命科学名著



人类分子遗传学（原书第三版）

Human Molecular Genetics (3th Edition)

〔英〕 T. 斯特罗恩
A .P. 里 德 编著

孙开来 主译

科学出版社

(Q-1819.31)

生命科学名著·典藏版



呈现国际学术精华，引进国际经典名著

细胞生物学精要（原书第三版）
分子生物学（原书第五版）
基因的分子生物学（原书第七版）
Brock微生物生物学（上下册）
遗传学：基因和基因组分析（第八版）
细菌分子遗传学（原书第五版）
神经生物学：从神经元到脑（原书第5版）
人类分子遗传学
衰老分子生物学
发育的原理（影印版）
糖生物学概述（原书第三版）
结构生物学：从原子到生命
Brock微生物生物学（影印）（原书第十版）
生物化学——基础理论与临床（原书第六版）
细胞生理学手册——膜生物物理学精要（下册）（原书第四版）
细胞生理学手册——膜生物物理学精要（上册）（原书第四版）
植物生物学
表观遗传学
蛋白质物理
进化
细胞
藻类学
基因组3
森林遗传学
基因组学概论
癌生物学
Lewin 基因X
糖生物学基础
微生物基因组
兽医微生物学



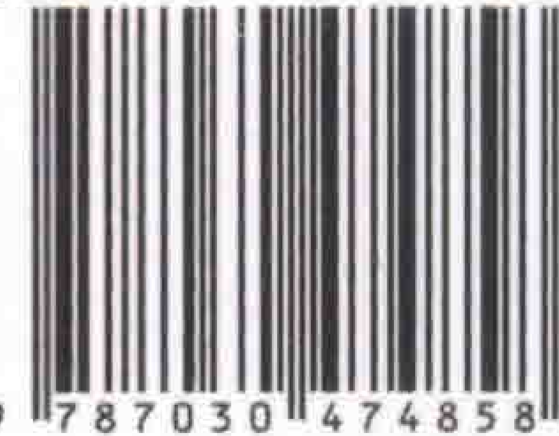
科学出版中心　生物分社
联系电话：010-64012501
E-mail：lifescience@mail.sciencep.com
网址：http://www.lifescience.com.cn

销售分类建议：生物学



赛拉艾芙
生命科学订阅号

www.sciencep.com


ISBN 978-7-03-047485-8
9 787030 474858 >

定价（全套）：4500.00元

生命科学名著·典藏版

人类分子遗传学

（原书第三版）

〔英〕T. 斯特罗恩　A.P. 里德　编著
孙开来　主译

科学出版社
北　京

图字：01-2005-6488 号

内 容 简 介

本系列丛书均选择生命科学领域经久不衰的经典名著，作者均为国际一流专家，堪称各个专业领域的国际第一书。每一本书的更新都紧跟学科发展，更加适合当前的学习和研究。

本丛书包括《癌生物学》、《分子生物学》、《神经生物学：从神经元到脑》、《表观遗传学》、《基因的分子生物学（第七版）》、《细胞生物学精要（原书第三版）》、《结构生物学：从原子到生命》等30本经典著作。

Human Molecular Genetics, 3rd, ed.
Tom Strachan, Andrew P . Read

图书在版编目（CIP）数据

生命科学名著：典藏版/（美）温伯格（Weinberg, R. A.）等编著；詹启敏等译. —北京：科学出版社, 2016
ISBN 978-7-03-047485-8

Ⅰ.①生… Ⅱ.①温…②詹… Ⅲ. ①生命科学–研究 Ⅳ.①Q1-0

中国版本图书馆 CIP 数据核字(2016)第 043879 号

责任编辑：王 静 李 悦
责任印制：张 伟 / 封面设计：刘新新

科学出版社 出版
北京东黄城根北街 16 号
邮政编码: 100717
http://www.sciencep.com
北京厚诚则铭印刷科技有限公司 印刷
科学出版社发行 各地新华书店经销
*
2016 年 7 月第 一 版 开本：787×1092 1/16
2016 年 7 月第一次印刷 印张：1431 3/4
字数：33 950 000
定价：4500.00 元
(如有印装质量问题，我社负责调换)

译者名单

主　译：孙开来

副主译：李　岭　邱广蓉　赵彦艳　富伟能

译　者：（按汉语拼音顺序）

陈芳杰　付　浩　富伟能　宫立国

贺　光　李春义　李福才　李　岭

李婷婷　李晓明　李英慧　刘　洪

刘丽英　吕晶玉　邱广斌　邱广蓉

尚　超　孙开来　王莉莉　赵彦艳

郑志红　周助人

谨以此书纪念

Frank Strachan（1921～2004）

中译本序

我们高兴地迎来中文版《人类分子遗传学》的问世。我们对许多能够轻易阅读英语文献的中国科学家充满敬意——但我们希望此中文版可使更多的学生、临床医师和年轻研究者能够无需纠缠于外文即能专注于其内容。

自本书第一版问世至今已有十年。这些年来，人类分子遗传学的世界并未停止带给我们惊讶与兴奋。“完成版”的人类基因组序列在2003年的公布标志着我们学科历史中一个重要时代的结束和另一个时代的开始。遗传学已经变得富含信息并且极其错综复杂。十年之前，研究者主要在研究单个基因。典型的研究方案将是致力于克隆某个基因，发现其功能并详细阐释该基因内突变所造成的后果。而今大部分研究则运用微阵列以及生物信息学来同时获取和分析大量基因的信息。研究者专注于分子通路而不是单个基因的产物，并试图运用系统生物学手段来阐明细胞的功能。表达微阵列允许经一次实验即可对大量基因的表达模式进行分析。DNA测序的价格每年都在下降，而测序的通量则在上升。刚问世的新型大规模平行测序技术将加快这一过程。甚至基因组分析都在从单一物种的分析进展到对基因组序列已知的许多物种的平行研究。所有这些进展正在共同实现对于我们的进化、发育以及基础生物学大量的新认识。

在人类疾病的研究中，这种重点变化最突出的例子在于癌生物学。诸如癌基因组解剖学计划（Cancer Genome Anatomy Project）之类的大型计划以及像基于微阵列的比较基因组杂交（Array-based Comparative Genomic Hybridisation）的全基因组技术正在为一个正常细胞转变为一个肿瘤原始细胞所发的改变提供新水平的深刻理解。基因组而非基因水平的研究是否将对其他疾病产生重要影响似乎尚不确定。事实上，迄今为止具有临床价值的进展几乎全部来源于对孟德尔疾病的研究。为寻找诸如多发性硬化、精神分裂症或者自闭症之类的常见疾病的易感基因所付出的巨大努力至今仅取得了令人失望的结果。但这种情况可能即将改变。超高精度的单核苷酸多态芯片（SNP chip）将使全基因组范围的关联分析成为可能，而单体型作图计划（HapMap project）则为了解我们的基因组如何由各种原始的染色体片段拼接而来提供了一个框架。然而，基因组研究所取得的进展能在多大程度上将医学由“诊断与治疗”转变到“预测与预防”的模式仍是一个深具争议的问题。

本书专注于原理而非大量的实例。我们并未准备为人类基因组或者遗传病提供一本详尽的事实说明。任何一个拥有个人台式电脑和快速互联网连接的人都能获取针对几乎任何遗传学题目的大量原始数据。然而要懂得这些数据的涵义则需要掌握遗传学的原理。我们尽力地解释了在人类基因组、其演化史和功能的现有见解之后的思路。我们希望读者将和我们一样发现这门学科是令人着迷的。

T. 斯特罗恩

A. P. 里德

2006年12日

Preface of Chinese Translation

We are delighted to welcome this Chinese Language edition of *Human Molecular Genetics*. We are full of admiration for the many Chinese scientists who are comfortable reading the literature in English. But we hope that this edition will allow many more students, clinicians and young researchers to concentrate on the content without having to wrestle with a foreign language.

It is now ten years since the first edition of our book appeared. The world of human molecular genetics has not ceased to surprise and excite us during those years. Publication of the "finished" human genome sequence in 2003 marked the end of one important era in the history of our subject, but the beginning of another. Genetics has become information-rich and massively parallel. Ten years ago researchers mainly investigated individual genes. A typical research project would aim to clone a gene, identify its function and detail the consequences of mutations in that gene. Now much research uses microarrays and bioinformatics to generate and process information simultaneously on large numbers of genes. Researchers focus on pathways rather than on individual gene products, and attempts are being made to interpret cellular processes using a systems biology approach. Expression microarrays allow the expression patterns of large numbers of genes to be analysed in a single experiment. Each year the cost of DNA sequencing falls and the throughput of sequences rises. New massively parallel sequencing technologies are just becoming available that will accelerate this process. Even genomic analysis is moving from analysis of a single species to parallel investigations of the many species whose genome sequences are now available. All these advances are combining to produce a flood of new understanding of our evolution, development and basic biology.

In studies of human disease, the most striking example of this change of emphasis is in cancer biology. Large-scale projects such as the Cancer Genome Anatomy project and whole-genome techniques like array-based Comparative Genomic Hybridisation, are providing a new level of insight into the changes that convert a normal cell into a tumour progenitor. Whether genomic, as opposed to genetic, research will have a major impact on other diseases is still somewhat uncertain. Up to now it remains true that clinically useful advances have almost all come from studies of mendelian conditions. Massive efforts to identify the genetic factors underlying susceptibility to common diseases such as multiple sclerosis, schizophrenia or autism have so far yielded disappointing results. Maybe this is about to change. Ultra-high resolution SNP chips allow whole genome association studies, and the HapMap project has provided a framework for understanding how our genome is made up of a mosaic of ancestral chromosome segments.

Nevertheless it remains a deeply controversial question how far advances in genomics can move medicine from a "diagnose and treat" to a "predict and prevent" model.

This book is about principles rather than facts. We do not aim to present a detailed factual account of either the human genome or genetic disease. Anybody with an ordinary PC and a fast internet connection can access vast volumes of raw data on almost any genetic topic. But understanding what the data mean requires an understanding of genetic principles. We have tried to explain the ideas behind current views of the human genome, its veolution and function, and the techniques used to develop these views. We hope readers find the subject as fascinating as we do.

Tom Strachan
Andrew P. Read
December 2006

译者序

《人类分子遗传学》(*Human Molecular Genetics*)一书的中译本现在出版了。此书是阐述人类遗传学的优秀教材。从它的第一版问世(1996年),我们就将其作为研究生教学的主要教材。依据人类基因组计划基本完成后所取得的众多发现,原书第三版(2004出版)作了许多新的补充和修订。

本书与同类教材比较有如下主要的优点:

第一,它较全面地涵盖了人类遗传学领域的基本内容。

第二,从人类基因组的DNA、RNA、蛋白质和基因以及基因组、细胞与个体的不同水平,系统而深入地阐述了遗传学原理。读者会从中学习到在一般同类书中难以得到的科学知识。

第三,它将遗传学原理和实验室应用的技术紧密结合,这对教学乃至研究是一本很好的指导书。

第四,它对各种类型的人类疾病,以典型的病例,结合环境因素与遗传因素,应用人类基因组的最新成就,阐述发病的机理,并展望人类基因组医学的发展前景。

第五,该书的一个特色是每章都插有全新的彩色图谱,非常有利于对理论和技术原理的理解。实际上,近年来国内有关人类和医学遗传学的教材都引用该书的内容和插图。中译本的插图,都为黑白的,但书后附有该书全书彩色插图的光盘。

因此,我相信该书的中译本出版对我国的人类医学遗传学的教学和研究工作的发展会起到重要促进作用。

中译本的出版是中国医科大学医学遗传学教研室全体同志在繁忙的教学和科研工作中共同努力的成果。同时得到科学出版社生物编辑部同志们的热情支持,在此对他们付出的辛勤劳动表示衷心地感谢!

希望读者从书中得到收益,并真诚盼望遗传学界对中译文提出批评与建议。

沈阳

2006年11月30日

前　　言

《人类分子遗传学》(原书第三版) 一书根据人类基因组计划后的众多发现做了最新的修订。进入后基因组时代，我们仍然确信此书能充当基础教材与研究文献之间的桥梁，以至于使相对缺乏该学科背景的人能了解和阅读最新的研究成果。

人类分子遗传学是一门大的学科。我们尽量将此书编成清楚区分色彩标记的几部分*，使用陈述性标题并以黑体字标示重要的新词汇，使之更易消化吸收。

本书的第一部分 (1～7 章) 涵盖了关于 DNA 结构和功能，染色体，细胞和发育，系谱分析的基本内容以及实验室所应用的基本技术。第二部分 (8～12 章) 讨论了各种基因组的测序计划，并洞察人类基因组的组成、表达、变异和进化。第三部分 (13～18 章) 的重点是对孟德尔疾病，复杂疾病和癌症的遗传起因的基因定位、鉴定和诊断。最后，第四部分 (19～21 章) 着眼于更广范围的功能基因组学、蛋白质组学、生物信息学、动物模型和治疗学。

我们提供了一个广泛的词汇表，并在第 5、6 和 12 章中附加三个特殊词汇表。还有两个索引：一个主要的索引和一个疾病索引**。

《人类分子遗传学》第二版发行以来的 4 年是一重要的时期，2001 年出版了人类基因组序列的草图，2003 年发表了它“完成的”版本。熟悉本书前一版的读者会注意到其中有许多改变：新增了关于细胞与发育和功能基因组学两章内容。关于复杂疾病部分，同基因组计划那章一样，已经完全重写和改编。在分子系统发生学 12 章新增的一节和介绍讨论含有新知识的伦理学图框中，是有许多小的改动。此外，考虑到过去 4 年的惊人进展，实际上每一页都有新的修改。令人赏心悦目的全彩色图谱意味着修改了所有的插图，其中许多是全新的*。如上所述，这些图 (除了引自某些其他刊物外) 都可从出版社的网站下载。

并非所有内容都做了改动。我们的目的仍是阐述原理，而非提供大量的事实。事实是易于得到的，主要是通过因特网，同时我们提供必要的参考索引。科学的规则要求，从研究水平综述来运用参考文献以肯定做出最初贡献的人们。然而，在本书每章末尾的进一步阅读书目具有更多教育的目的；所以我们通常引用的是综述而不是有关一个主题的第一篇论文，并从易于找到的杂志选择参考文献。因而我们希望没有找到参考他们创新论文的人们会予以理解。如同在前一版中一样，我们想要传达快速进展研究的感受，并希望读者为了人类基因组中的连续进程的发现，有朝一日会与我们分享快乐和热情。

同往常一样，我们感谢对本书前一版的许多评论者，即使我们未能总是吸收他们的建议和意见。对于这次新版，我们感谢许多同事对各章的意见和评述，特别应提到 Gavin Cuthbert，Ian Hampson，Mike Jackson，Ralf Kist，Chris Mathew，Heiko Peters，

---

* 本书为黑白印刷，全书彩图制成光盘，放在书后。——出版者

** 原书索引均未保留，该翻译书重新生成索引。——出版者

Nalin Thakker，Andy Wallace 和 John Wolstenholme。应特别感谢 Richard Twyman 对 3、19 和 20 章的重要帮助。最高兴的是一直与 Jonathan Ray、Fran Kingston 和 Garland Science/BIOS 科学出版社的全体人员以及与开发全彩色图的 Touchmedia 一起工作。最后，对与我们长期患难与共的家庭成员特别是 Meryl、Alex、James 和 Gilly，以及秘书 Anne、Kate、Leanne 和 Margaret，对他（她）们的支持和付出全部的努力所给予的关怀表示谢意。

Tom Strachan and Andrew Read
汤姆·斯特罗恩和安德鲁·里德

（孙开来 译）

补充学习帮助

教授们可得到的补充材料包括：

人类分子遗传学 3 的艺术品

出于展示的目的，含有来自此书的所有图片的 CD-ROM。图片适用于 JEPG 和 PowerPoint 格式，对那些采用正文的演讲者是免费的。也可购买获得。

Garland Science Classwire ™

我们乐意为《人类分子遗传学 3》的采用者提供 Garland Science Classwire™，此网址允许你：

▶ 获得由 Garland Science 提供的教学资源

▶ 在几分钟内为你的课程创造一个定制的网址

▶ 与你的学生在线交流

▶ 建立一个持续扩大的教学资源文库

对于《人类分子遗传学 3》，在线教学资源包括：

▶ 书中所有的图像，存在于一个能够下载的、网上可立即得到的格式或 PowerPoint 可立即得到的格式。

▶ 进入 Classwire 所拥有的所有其他 Garland Science 资源的通路。

《人类分子遗传学 3》的采用者无限利用 Garland Science Classwire 服务。你可以在 www. classwire. com/garlandscience/demol. html. 察看 Classwire™ 的一个在线说明。

至于更多的细节，请与你本地销售商代表联系。

Classwire 是 Chalkfree 公司的一个商标。

在我们开始阅读之前智能使用 Internet

无需告诉现在的学生和研究者去使用 Internet。然而，有几点特别与本书的读者相关，我们想在开始的时候如实地指出。

在这本书中，我们试图涵盖人类分子遗传学的原理，但是我们无法列出许多论据。当我们给出论据时，它们主要在那里解释原理。但是没有论据的原理相当无效，我们希望你们必要时从 Internet 查询论据。基因组计划，以及所有人类遗传学的相关研究已经产生了一个真正的数据浪潮。聪明且有辨别力的使用 Internet 是任何科学家的一个关键技能，但也许尤其对于遗传学家和遗传学学生（更关键）。

我们仅仅运行 Google 检索“遗传学”，它产生了 3 630 000 个匹配查询结果。那些网址中有一些是重要的资源，许多是次要的，一些则是蓄意骗人的、不精确的。在这本书中，我们为特定的主题推荐了许多网址。我们建议下面核心网址；我们选择它们，因为它们是可信的、稳定的（在本书的生存期内不可能改变），并提供了与许多其他网址非常有效的链接。以此作为起点，你应该能够逐渐形成你自己的有用的网址表并从Internet上惊人的财富中明显受益。

▶ 遗传数据的普遍起点：http://www. ncbi. nlm. nih. gov；http://www. ebi. ac. uk/services/

▶ 对于基因组数据：www. ensembl. org；http://genome. cse. ucsc. edu

▶ 对于蛋白质信息：http://ca. expasy. org；

▶ 对于任何孟德尔表型信息：http://www. ncbi. nim. nih. gov/omim/

▶ 进入生物医学文献的通路：http://www. ncbi. nlm. nih. gov/entrez/

遗传密码（不同的线粒体密码见图 1.22）

第二个碱基	第一个碱基			
	U	C	A	G
U	UUU } Phe UUC } UUA } Leu UUG }	CUU } CUC } Leu CUA } CUG }	AUU } AUC } Ile AUA } AUG Met	GUU } GUC } Val GUA } GUG }
C	UCU } UCC } Ser UCA } UCG }	CCU } CCC } Pro CCA } CCG }	ACU } ACC } Thr ACA } ACG }	GCU } GCC } Ala GCA } GCG }
A	UAU } Tyr UAC } UAA STOP UAG STOP	CAU } His CAC } CAA } Gln CAG }	AAU } Asn AAC } AAA } Lys AAG }	GAU } Asp GAC } GAA } Glu GAG }
G	UGU } Cys UGC } UGA STOP UGG Trp	CGU } CGC } Arg CGA } CGG }	AGU } Ser AGC } AGA } Arg AGG }	GGU } GGC } Gly GGA } GGG }

一个字母的氨基酸代码

A 丙氨酸	C 半胱氨酸	D 天冬氨酸	E 谷氨酸	F 苯丙氨酸
G 甘氨酸	H 组氨酸	I 异亮氨酸	K 赖氨酸	L 亮氨酸
M 蛋氨酸	N 天冬酰胺	P 脯氨酸	Q 谷氨酰胺	R 精氨酸
S 丝氨酸	T 苏氨酸	V 缬氨酸	W 色氨酸	Y 酪氨酸
X 指一个终止密码				

补充信息

下列主题的信息可在提示页中找到。

电子数据库和资源 URL（又见“在我们开始阅读之前智能使用 Internet”）

缩 略 语

缩略语	英 文	中 文
2D	two-dimensional	二维
2DGE	two-dimensional gel electrophoresis	二维凝胶电泳
3′UTR	3′Untranslated region	3′非翻译区
5′UTR	5′ Untranslated region	5′非翻译区
5-MeC	5-Methyl cytosine	5-甲基胞嘧啶
A	adenine	腺嘌呤
AAV	adeno-associated virus	腺病毒相关病毒
AcMNPV	autographa californica nuclear polyhedrosis virus	苜蓿银纹夜蛾（苜蓿尺蠖）核型多角体病毒
ADAR	adenosine deaminase acting on RNA	双链 RNA 特异性腺苷脱氨酶
AID	activation-induced deaminase	激活诱导的脱氨酶
ALL	acute lymphoblastoid leukemia	急性淋巴细胞性白血病
AMH	anti-Mullerian hormone	抗中肾旁管激素（抗苗勒氏管激素）
AML	acute myeloid leukemia	急性粒细胞性白血病
ARF	alternative Reading Frame	选择性读框
ARMS	amplification Refractory Mutation System	扩增受阻突变系统
ARS	autonomously replicating sequence	自主复制序列
AS	Angelman syndrome	Angelman 综合征
ASO	allele-specific oligonucleotide	等位基因特异的寡核苷酸
ASP	affected sib pair	受累同胞对
AT	ataxia telangiectasia	共济失调性毛细血管扩张症
ATCC	American Type Culture Collection	美国标准菌库
AVE	anterior visceral endoderm	前脏壁内胚层
BAC	bacterial artificial chromosome	细菌人工染色体
BCR	breakpoint cluster region	断裂点簇集区
BER	base excision repair	碱基切除修复
BMI	body mass index	体重指数
BOR	branchio-oto-renal syndrome	腮-耳-肾综合征

续表

缩略语	英　文	中　文
bp	base pairs	碱基对
BrdU	bromodeoxyuridine	溴脱氧尿苷
C	cytosine	胞嘧啶
CATH	Class Architecture Topology Homologous superfamily	等级结构拓扑学同源超家族（蛋白质结构分类数据库之一）
CCC	covalently closed circular	共价闭合环状
CCM	chemical cleavage of mismatch	错配的化学裂解
CD	crohn's disease	克隆氏病
cDNA	complementary DNA	互补 DNA
CDR	complementarity-determining region	互补性决定区
CEN	centromere element	着丝粒元件
CF	cystic fibrosis	囊性纤维化
CGH	comparative genomic hybridization	比较基因组杂交
CID	chemically-induced dimerization	化学诱导二聚化
CIN	chromosomal instability	染色体不稳定
cM	centiMorgan	厘摩
CNS	central nervous system	中枢神经系统
C_0t	product of DNA concentration and time	DNA 浓度与时间的产物
CREB	CRE-binding protein	CRE 结合蛋白
CS	cockayne syndrome	科凯恩综合征
D	displacement or diversity	置换或多样性
ddNTP	dideoxynucleoside triphosphate	双脱氧核苷三磷酸
DGGE	denaturing gradient gel electrophoresis	变性梯度凝胶电泳
dHPLC	denaturing high performance liquid chromatography	变性高效液相色谱
DIGE	difference gel electrophoresis	差异凝胶电泳
DMD	duchenne muscular dystrophy	进行性假肥大性肌营养不良（杜兴肌营养不良）
DNA	deoxyribonucleic acid	脱氧核糖核酸
DnasI	deoxyribonuclease I	脱氧核糖核酸酶 I
DOP-PCR	degenerate oligonucleotide primed polymerase chain reaction	简并寡核苷酸引物聚合酶链反应
ds	double-stranded	双链的

续表

缩略语	英 文	中 文
DS	down syndrome	Down 综合征（唐氏综合征；先天愚型）
DZ	dizygotic	双卵的；异卵的
EBV	Epstein-Barr virus	EB 病毒（埃巴病毒，非洲淋巴瘤病毒）
EC	embryonal carcinoma	胚胎性癌
ECACC	European Collection of Cell Cultures	欧洲细胞培养库
ECM	extracellular matrix	细胞外基质
EG	embryonic germ	胚胎生殖细胞
EGF	epidermal growth factor	表皮生长因子
ELSI	ethical legal and societal implications	伦理、法律和社会影响
EMS	ethyl methylsulfonate	乙烷基甲基磺胺
ENU	ethyl nitrosurea	乙烷基硝基尿素
ER	endoplasmic reticulum	内质网
ERCC	excision repair cross-complementing	切除修复交叉互补
ERV	endogenous retroviral sequence	内源性反转录病毒序列
ES	embryonic stem	胚胎干细胞
ESE	exonic splice enhancer	外显子剪接增强子
ESI	electrospray ionization	电喷射离子化（作用）
ESS	exonic splice silencer	外显子剪接沉默子
EST	expressed sequence tag	表达序列标签
EtBr	ethidium bromide	溴化乙锭
ETDT	extended TDT	扩展 TDT
FAP	familial adenomatous polyposis	家族性腺瘤性息肉病
FISH	fluorescence *in situ* hybridization	荧光原位杂交
FITC	fluorescein isothocyanate	异硫氯酸荧光素
FLAM	free left Alu monomer	游离的左侧 Alu 单体
FRAM	free right Alu monomer	游离的右侧 Alu 单体
FSSP	fold classification based on Structure-Structure alignment of Protein	基于蛋白质结构比对的折叠分类法
G	guanine	鸟嘌呤
gcv	ganciclovir	羟甲基无环鸟苷（甘昔洛韦）
GDB	genome database	基因组数据库

续表

缩略语	英　文	中　文
GE	genome equivalents	基因组当量
GFP	green fluorescent protein	绿色荧光蛋白
GO	gene Ontology	基因本体论
GSS	gerstmann-Straussler-Scheinker	GSS 病
GST	glutathione-S-transferase	谷胱甘肽-S-转移酶
H	heavy	重
HAT	histone acetyltransferase	组蛋白乙酰转移酶
HD	huntington disease	亨廷顿病
HDAC	histone deacetylase	组蛋白脱乙酰基酶
HERV	human endogenous retroviral sequence	人类内源性反转录病毒序列
HGDP	Human Genome Diversity Project	人类基因组多样性计划
HGP	Human Genome Project	人类基因组计划
HGT	horizontal gene transfer	水平基因转移
HLA	human leukocyte antigen	人类白细胞抗原
HLH	helix-loop-helix	螺旋-环-螺旋
HNPCC	hereditary nonpolyposis colon cancer	遗传性非息肉性结肠癌
HPLC	high pressure liquid chromatography	高压液相色谱
HPRT	hypoxanthine guanine phosphoribosyl transferase	次黄嘌呤-鸟嘌呤磷酸核糖转移酶
HSC	hematopoetic stem cell	造血干细胞
HSCR	hirschsprung disease	先天性巨结肠症（希施斯普龙病）
HSV-TK	herpes simplex virus thymidine kinase	单纯疱疹病毒胸苷激酶
HTH	helix-turn-helix	螺旋-转角-螺旋
HUGO	Human Genome Organization	人类基因组机构
HDV	human delta virus	人类 δ 病毒
IBD	identity by descent or Inflammatory bowel disease	血源同一或炎性肠病（肠炎）
IBS	identity by state	状态同一
ICAT	isotope coded affinity tag	同位素编码亲和标签
ICLC	interlab Cell Line Collection	实验室间的细胞系库
ICM	inner cell mass	内细胞团
ICSI	intracytoplasmic sperm injection	精子卵浆内注射技术

续表

缩略语	英　文	中　文
Ig	immunoglobulin	免疫球蛋白
IM	intermediate mesoderm	间介中胚层
IP_3	inositol 1，4，5-trisphosphate	1，4，5-三磷酸肌醇
IPG	immobilized pH gradient	固相 pH 梯度
IPTG	isopropyl-thio-β-D-galactopyranoside	异丙基-β-D-硫代半乳糖苷
IRE	iron-response element	铁反应元件
ISCN	International System for Human Cytogenetic Nomenclature	国际人类细胞遗传学术语命名法
IVF	*In vitro* fertilization	体外受精
J	joining	连接术
kb	kilobases	千碱基
KEGG	Kyoto Encyclopedia of Genes and Genomes	基因和基因组的京都百科全书
L	light	轻
LCR	locus control region	基因座控制区
LD	linkage disequilibrium	连锁不平衡
LINES	long interspersed nuclear element	长散在核元件
LoH	loss of heterozygosity	杂合性丢失
LPM	lateral plate mesoderm	侧板中胚层
LTR	long terminal repeat	长末端重复
m^7G	7-Methylguanosine	7-甲基鸟苷
mAb	monoclonal antibody	单克隆抗体
MAD	multi-wavelength anomalous dispersion	多波长反常色散
MALDI-TOF	matrix-assisted laser desorption/ionization time-of-flight	基质辅助激光解吸/电离飞行时间
MS	mass spectrometry	质谱测定法；质谱分析法；质谱法
MAPH	multiplex amplifiable probe hybridization	多重可扩增探针杂交技术
MAR	matrix attachment region	核基质附着区
Mb	mega base	兆碱基
MCS	multiple cloning site	多克隆位点
M-FISH	multiplex FISH	复合 FISH
MGSC	Mouse Genome Sequencing Consortium	国际小鼠基因组测序合作组织
MIN	microsatellite instability	微卫星不稳定性

续表

缩略语	英　文	中　文
MIR	mammalian-wide interspersed repeat	哺乳类散在重复
miRNAs	microRNA	微 RNA
MM	mismatch	错配
MODY	maturity onset diabetes of the young	青春晚期糖尿病
mRNA	messenger RNA	信使 RNA
MS	mass spectrometry	质谱测定法；质谱分析法；质谱法
MS/ MS	tandem mass spectroscopy	串联质谱
mtDNA	mitochondrial DNA	线粒体 DNA
MTOC	microtubule-organizing center	微小管形成中心
MYr	million years	百万年
MZ	monozygotic	同卵的；单卵的
NAS	nonsense-associated altered splicing	无义密码子相关的可变剪接
NBS	Nijmegen breakage syndrome	Nijmegen 染色体断裂综合征
NCAM	neural cell adhesion molecule	神经细胞黏附分子
NER	nucleotide excision repair	核苷酸切除修复
NFI	neurofibromatosis type I	神经纤维瘤 I 型
NMR	nuclear magnetic resonance	核磁共振
NOE	nuclear Overhauser effect	核欧沃豪斯效应
NPL	nonparametric lod	非参数对数优势比
NRSE	neural restrictive silencer element	神经限制性沉默元件
NRSF	neural restrictive silencer factor	神经限制性沉默因子
OD	optical density	光密度；吸光度
OI	osteogenesis imperfecta	成骨不全
OLA	oligonucleotide ligation assay	寡核苷酸连接分析法
PAC	P1 artificial chromosome	P1 噬菌体人工染色体
PCNA	proliferating cell nuclear antigen	增殖细胞核抗原
PCR	polymerase chain reaction	聚合酶链反应
PDB	protein Databank	蛋白质数据库
PEG	polyethylene glycol	聚乙二醇
PFD	polyostotic fibrous dysplasia	多骨纤维性发育不良
PFGE	pulsed field gel electrophoresis	脉冲场凝胶电泳

续表

缩略语	英 文	中 文
PGC	primordial germ cell	原始生殖细胞
Ph	philadelphia	费城
PIC	polymorphism information content	多态信息含量
PIP	phosphatidyl inositol 4，5-bisphosphate	4，5二磷酸磷脂酰肌醇
PKD1	adult polycystic kidney disease	成人多囊肾疾病
PKU	phenylketonuria	苯丙酮尿症
PM	perfect match or Paraxial mesoderm	完全匹配或轴旁中胚层
PMF	peptide mass fingerprinting	肽质量指纹图谱
PML	promyelocytic leukemia	前髓细胞性白血病
PP	pyrophosphate residue	焦磷酸盐残基
PSI-BLAST	position-specific iterated BLAST	位置特异性叠代 BLAST
PTT	protein truncation test	蛋白质截断实验
Pu	purine	嘌呤
PWS	Prader-Willi syndrome	肌张力减低-智力减低-性腺功能减退-肥胖综合征（普-韦二氏综合征）
Py	pyrimidine	嘧啶
QTL	quantitative trait locus	数量性状基因座
RACE	rapid amplification of cDNA end	cDNA 末端快速扩增法
REMI	restriction enzyme-mediated integration	限制性内切酶介导的整合
RER	rough endoplasmic Reticulum	粗面内质网
REST	RE-1 silencing transcription factor	RE-1 转录沉默因子
RF	replicative form	复制型
RFLP	restriction fragment length polymorphism	限制性片段长度多态性
RISC	RNA induced silencing complex	RNA 诱导沉默复合体
RMSD	root mean square deviation	标准差
RNA	ribonucleic acid	核糖核酸
RNAi	RNA interference	RNA 干扰
RNase	ribonuclease	核糖核酸酶
RNP	ribonucleoprotein	核糖核蛋白
rNTP	ribonucleoside triphosphate	核糖核苷三磷酸
RP	retinitis pigmentosa	视网膜色素变性
rRNA	ribosomal RNA	核糖体 RNA

续表

缩略语	英　文	中　文
RSP	restriction site polymorphism	限制位点多态性
RT	reverse transcriptase	反转录酶
RT-PCR	reverse transcriptase-polymerase chain reaction	反转录-聚合酶链反应
SA	splice acceptor	剪接受体
SAGE	serial analysis of gene expression	基因表达的系列分析
SAR	scaffold attachment regions	支架附着区
SCA1	spinocerebellar ataxia type 1	脊髓小脑性共济失调 1 型
scFv	single chain variable fragment	单链可变区片段
SCID	severe combined immunodeficiency	重度联合免疫缺陷症
SCOP	Structural Classification of Protein	蛋白质结构分类
SD	splice donor	剪接供体
SDS	sodium dodecyl sulfate	十二烷基硫酸钠
SF1＋	steroidogenic factor 1	类固醇转录因子 1
SINES	short interspersed nuclear element	短散在核元件
SIRAS	single isomorphous replacement with anomalous scattering	单一同型置换伴非寻常散射
siRNA	short interfering RNA	短干扰 RNA
snoRNA	small nucleolar RNA	核仁小 RNA
SNP	single nucleotide polymorphism	单核苷酸多态性
snRNA	small nuclear RNA	核内小 RNA
SCA	spino-cerebellar ataxia	脊髓小脑性共济失调
SRP	signal recognition particle	信号识别颗粒
ss	single-stranded	单链的
SSR	simple sequence repeats	简单重复序列
SSRP	simple sequence repeats polymorphism	简单重复序列多态性
STAT	signal transducers and activators of transcription	信息传递与转录激活因子
stRNA	small temporal RNA	小分子时序 RNA
STRP	short tandem repeat polymorphism	短串联重复序列多态性
STS	sequence tagged site	序列标签位点
SV40	simian virus 40	猿猴病毒 40；猿猴空泡病毒
SVAS	supravalvular aortic stenosis	主动脉瓣上狭窄

续表

缩略语	英 文	中 文
T	thymine	胸腺嘧啶
TCR	T-cell receptor	T 细胞受体
TCS	treacher Collins syndrome	下颌面骨发育障碍（特雷歇·柯林斯氏综合征）
TDT	transmission disequilibrium test	传递不平衡检验
TF	transcription factor	转录因子
TFR	transferrin receptor	转铁蛋白受体
TGF	transforming growth factor	转化生长因子
TIGR	the Institute for Genome Research	基因组研究中心
TK	thymidine kinase	胸苷激酶
Tm	melting temperature	解链温度；熔解温度
TMP	thymidine monophosphate	胸苷酸
TNF	tumor necrosis factor	肿瘤坏死因子
TOF	time of flight	飞行时间
TPA	tissue plasminogen activator	组织纤维蛋白溶酶原激活剂
tRNA	transfer RNA	转移 RNA
TS	tumor suppressor	肿瘤抑制基因
TTD	trichothiodystrophy	毛发低硫营养不良
U	uracil	尿嘧啶
UC	ulcerative colitis	溃疡性结肠炎
UEC	unequal crossover	不等交换
UESCE	unequal sister chromatid exchange	不等姐妹染色单体交换
UPD	uniparental disomy	单亲二倍体
UTR	untranslated region	非翻译区
UV	ultraviolet	紫外线
V	variable	可变的
VS	Varkud satellite	Varkud 卫星
VCFS	velocardiofacial syndrome	软腭-心-面综合征
VNTR	variable number tandem repeat	可变数目串联重复
VPC	vector-producing cell	载体生产细胞
VAGR	Wilms tumor aniridia genital abnormalities mental retardation	肾母细胞瘤（Wilms 瘤）伴无虹膜、生殖器异常、智力低下

续表

缩略语	英　文	中　文
VLS	Williams syndrome	主动脉瓣上狭窄症候群（威廉斯综合征）
WS1	waardenburg syndrome type 1	瓦尔敦堡综合征 1 型
Xgal	5-bromo 4-chloro 3-indolyl β-*D*-galactopyranoside	5-溴-4-氯-3-吲哚-β-D-乳糖苷
XP	xeroderma pigmentosum	着色性干皮病
X-SCID	X-linked severe combined immunodeficiency disease	X-连锁重度联合免疫缺陷症
YAC	yeast artificial chromosome	酵母人工染色体
ZPA	zone of polarizing activity	极性活性区

目　　录

第1章　DNA结构和基因表达

本章内容

1.1　DNA，RNA及多肽中的构件和化学键

分子遗传学首要关心的是信息大分子DNA（**脱氧核糖核酸**，deoxyribonucleic acid）和RNA（**核糖核酸**，ribonucleic acid）间的相互关系，以及这些分子如何用于合成所有蛋白质的基本组成部分——**多肽**（polypeptide）。在某些病毒中，RNA是遗传物质，但是在所有细胞中，遗传信息储存于DNA分子中。细胞DNA分子被选择的区域作为合成RNA分子的模板。绝大多数RNA分子依次在基因表达的不同阶段直接或辅助地用于指导多肽的合成。因为绝大多数基因表达都参与多肽的合成，所以蛋白质体现了DNA的主要功能性终点因而是一个细胞净重的大部分。**蛋白质**（protein）术语源于希腊语“proteios”一词，意思为“第一等级的”，并反映了在各种各样的细胞功能中蛋白质的重要作用，它们可作为酶、受体、储存蛋白、转运蛋白、转录因子、信号分子和激素等而起作用。

1.1.1　DNA，RNA和多肽是由一简单重复单位的线性序列所限定的大分子多聚体

在真核细胞中，单个DNA分子存在于细胞核的染色体、线粒体中，也存在于植物细胞的叶绿体中。它们是大分子多聚体，具有一个由交替的糖和磷酸残基组成的线性骨架。DNA分子中的糖是**脱氧核糖**（deoxyribose），它是一种五碳糖，且连续的糖的残基通过共价的磷酸二酯键连接起来。一个含氮的碱基共价地连接到每个糖残基的第一个碳原子上。共发现四种碱基：**腺嘌呤**（adenine，A），**胞嘧啶**（cytosine，C），**鸟嘌呤**（guanine，G）和**胸腺嘧啶**（thymine，T），它们由碳原子和氮原子的杂环构成。碱基可以被分为两类：

▶ **嘌呤**（purine）（A 和 G）有两个闭合的杂环；

▶ **嘧啶**（pyrimidine）（C 和 T）有一个闭合的杂环。

一个糖与一个附带的碱基称为核苷。一个核苷与一个附加在 5′或者 3′碳原子的磷酸基构成一个**核苷酸**（nucleotide），它是一条 DNA 链的基本重复单位（图 1.1 和图 1.2）。RNA 分子组成与 DNA 分子组成相似，不同的是它们含有**核糖**（ribose）残基替代脱氧核糖以及**尿嘧啶**（uracil，U）替代胸腺嘧啶（图 1.1 和图 1.2）。



图 1.1　核酸及相应的核苷与核苷酸中常见的碱基

注：通常在描述一个具有单个单磷酸的核苷酸的名字时，用后缀-ylate 代替碱基的后缀-ine，就像腺苷酸、鸟苷酸等。TMP、TDP 和 TTP 周围的括号（[]）表明它们不常见。

蛋白质由一个或者更多的多肽分子构成，它们可通过添加各种各样的碳水化合物侧链或者其他化学基团而被修饰。与 DNA 和 RNA 一样，多肽分子是由一个重复单位的线性序列构成的多聚体，在这里，这些重复单位是**氨基酸**（amino acid）。后者由一个带有正电荷的氨基和一个带有负电荷的羧酸（羧基）通过一个附带一条可识别侧链的中心碳原子连接到一起而构成。20 种不同的氨基酸依据它们侧链的性质可分为不同的类型（图 1.3）。分类如下：

▶ **碱性**（basic）氨基酸含有一条带一个净正电荷的侧链；在生理性 pH 值下，侧链上的一个氨基（NH_2）或组氨酸环获得一个 H^+ 离子；

▶ **酸性**（acidic）氨基酸含有一条带一个净负电荷的侧链；在生理性 pH 值下，侧链上的一个羧基基团失去一个 H^+ 离子形成 COO^-；

▶ **无电荷极性**（uncharged polar）氨基酸是电中性的，但是含有极性电基团的侧链，这些电基团由于具有部分电荷（用 δ^+ 或 δ^- 来表示）而得以区分。例如，**羟基**（hydroxyl group）（—OH）和**巯基**（sulfhydryl group）（—SH）中的氢原子都带有部分正电荷，而氧/硫原子具有部分负电荷，导致如下的标示：（$—O^{\delta^-}—H^{\delta^+}$）和

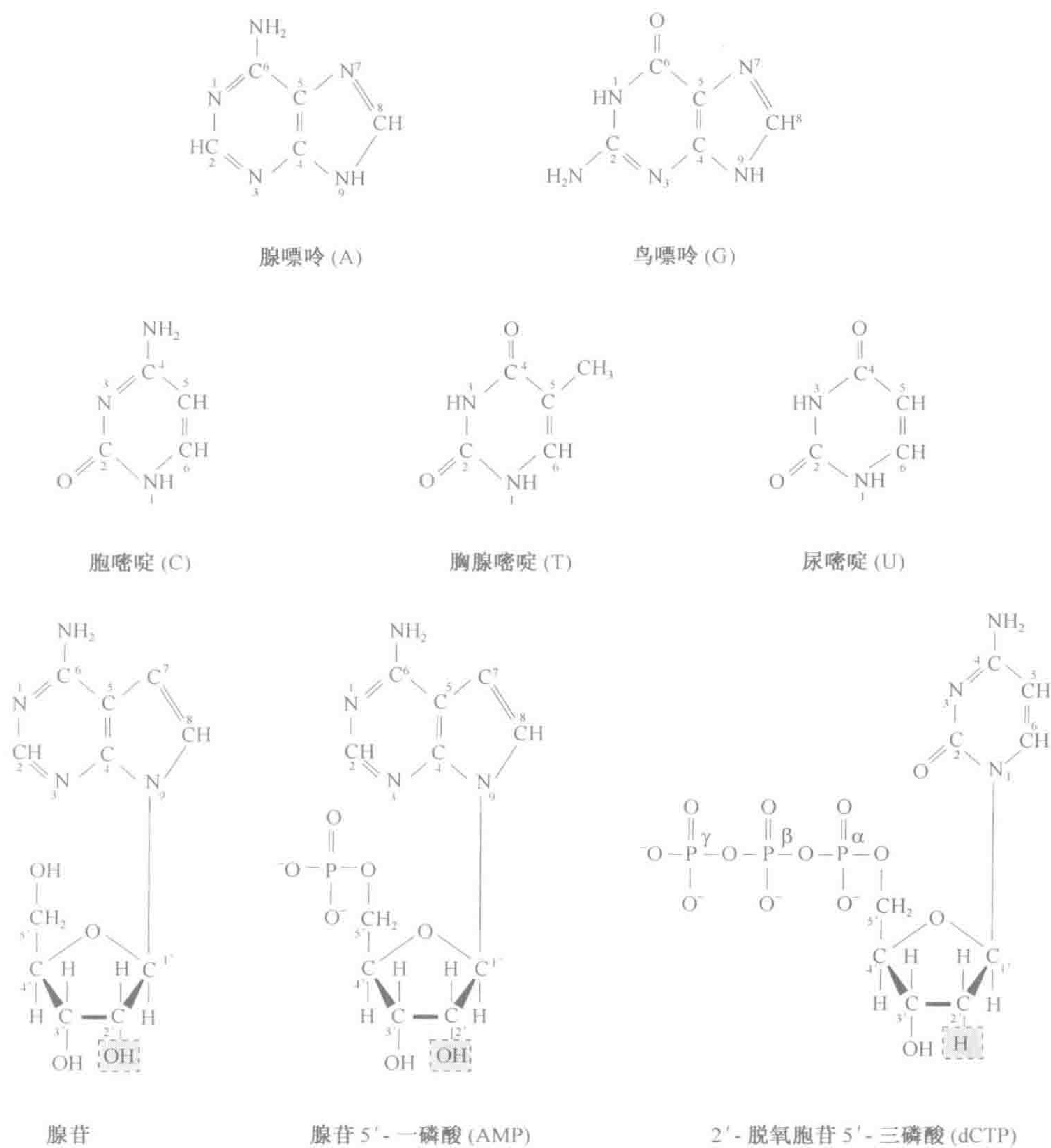


图 1.2　碱基、核苷及核苷酸的结构

糖环底部粗线用来表示环的平面对于相应的碱基平面而言固定于 90°角（即如果一个碱基的平面表示为平放于页的表面，那么糖的 2′和 3′碳原子可被认为是向上突出于页表面而氧原子是向下突出于页表面）。**注：** 脱氧核糖与核糖的编号限定于 5 个碳原子，命名为 1′到 5′，但是碱基的编号包括出现在杂环的碳原子和氧原子。与 2′碳原子连接的、突出显示的羟基和氢原子表示核糖与脱氧核糖残基间的本质区别。磷酸基团依据与糖环的远近连续地表示为 α、β、γ 等（见 dCTP 结构）。

$(-S^{\delta-}-H^{\delta+})$。

▶ **非极性中性**（nonpolar neutral）氨基酸是**疏水的**（hydrophobic）（排斥水的）。它们通常彼此相互作用或与其他疏水基团相互作用。

多肽是通过一个氨基酸的氨基团和相邻氨基酸的羧基之间的缩合反应形成的，缩合反应形成一个**重复骨架**（repeating backbone）（—NH—CHR—CO—），其中氨基酸的 R 侧链彼此不同（图 1.21）。

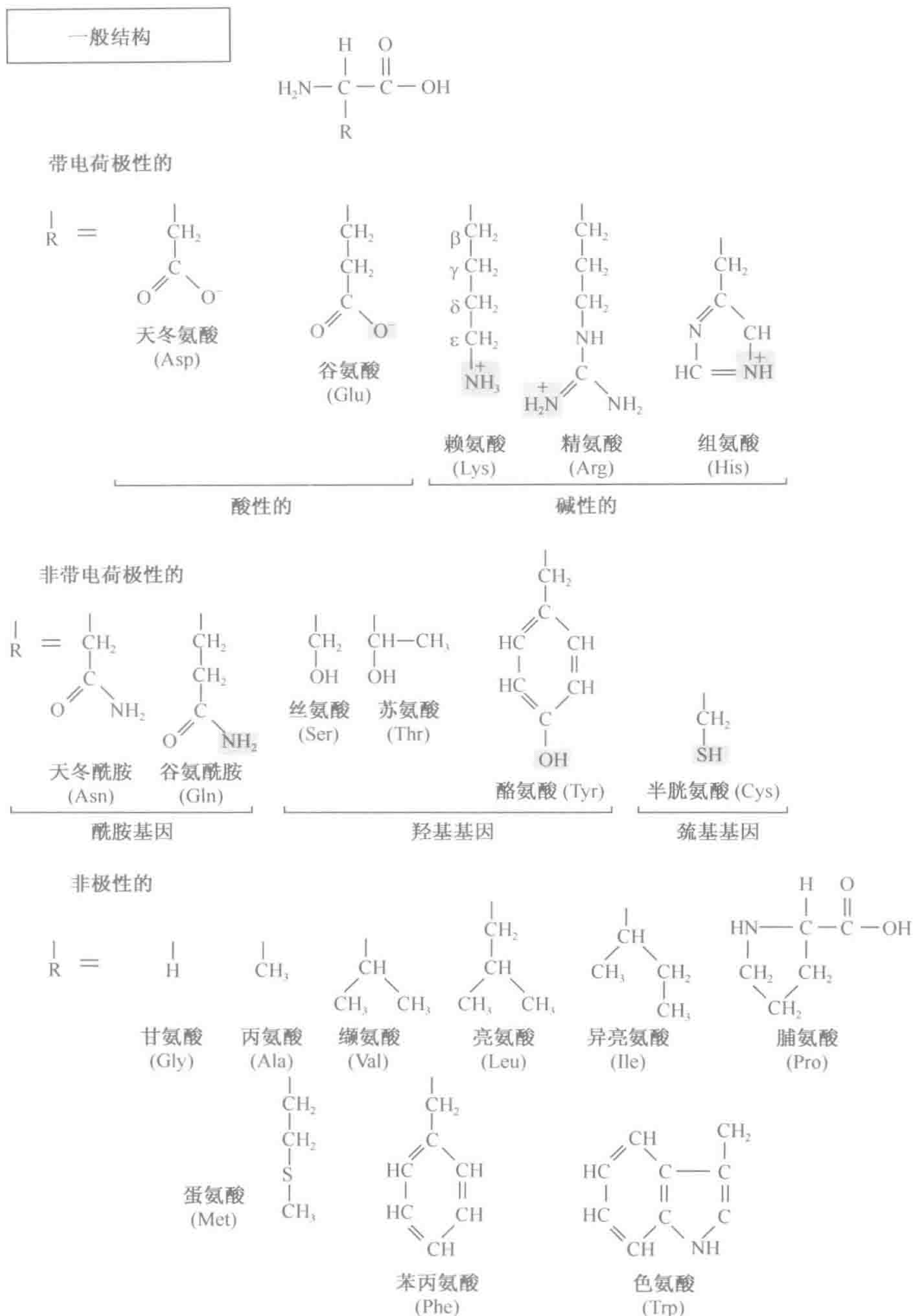


图 1.3　20 种主要氨基酸的结构

一个亚类（例如酸性氨基酸）的氨基酸在化学上非常相似。突出显示的基团是极性化学基团。编号碳原子的惯例命名中心碳原子为 α，而命名随后的线性侧链的碳原子为 β、γ、δ 等（见顶部赖氨酸侧链的例子）。一般来说，极性氨基酸是亲水的而非极性氨基酸是疏水的，但是甘氨酸（具有一条非常小的侧链）和半胱氨酸（其巯基极性不如一个羟基）在亲水-疏水等级中占据中间的位置。**注**：脯氨酸因为其侧链既与氨基的氮原子相连，也与中心碳原子相连而与众不同。

1.1.2　共价键赋予稳定性；较弱的非共价键促进分子间缔合作用并稳定结构

核酸和蛋白质多聚体的稳定性主要依赖于连接它们线性骨架组成原子的强**共价键**（covalent bond）。除了共价键之外，许多弱的**非共价键**（noncovalent bond）（表 1.1）对于这些分子之间以及一单个核酸或蛋白质分子内基团之间的相互作用很重要。一般来说，这样的非共价键要比共价键弱 10 倍以上。共价键的强度只是由相关的特定原子决定。与共价键不同，非共价键的强度关键依赖其水环境。由于单个水分子间发生非共价键形成的一个快速变化的网络，水的结构特别复杂，在此结构中主要的力是**氢键**（hydrogen bond），它是一种由一个部分带正电氢原子和一个部分带负电的原子间形成的弱静电键，对于水分子来说，部分带负电的原子是氧原子。

表 1.1　弱的非共价键形成

键的类型	键的性质
氢键	当氢原子被夹在两个电子吸引的原子（通常为氧原子或氮原子）之间时形成氢键。它们在核酸和蛋白质的结构与功能中的重要性的例子见框 1.1。
离子键	离子作用发生于带电基团之间。它们在石英晶体中可以非常强，但在水环境中，带电基团受到水分子或溶液中其他离子的保护，因而非常弱。不过它们在生物功能方面（例如对于酶-底物识别来说）非常重要。
范德华力键	任何两个彼此非常靠近的原子，由于它们波动的电荷而表现出一个弱吸引结合的相互作用（范德华力吸引），直到它们变得极度靠近时彼此非常强烈地排斥（范德华力排斥）。虽然单个的范德华力吸引非常弱，但是当两个大分子表面之间非常适合时，范德华力吸引可变得很重要。
疏水力键	水是一个极性分子。当疏水分子或化学基团位于含水环境时，为了使它们对水分子间复杂的氢键形成网络的破坏作用最小化，它们被强迫在一起。据说以这种方式被强迫在一起的疏水基团是通过疏水键维持在一起的，即使他们相互吸引的基础是由于水分子普通的排斥所致。

带电荷的分子在水中是高度可溶的。由于存在于 DNA 和 RNA 核苷酸组成成分中的磷酸电荷，所以 DNA 和 RNA 都是带有负电的（多聚阴离子）。依赖于它们的氨基酸组分，蛋白质可以带有一个净正电荷（**碱性蛋白质**，basic protein）或一个净负电荷（**酸性蛋白质**，acidic protein）。水分子氢键形成的电位意味着带有极性基团的分子（包括 DNA、RNA 和蛋白质）能够与水分子形成多重相互作用，导致它们溶解。因此，如果它们含有明显数目的带电的或者中性的极性氨基酸，那么甚至电中性的蛋白质也常常易于溶解。相反，膜结合蛋白常以高含量的疏水性氨基酸为特征，这些氨基酸在一个脂膜的疏水环境中热动力更稳定。

共价键需要相当大的能量输入才能破坏，与此不同，非共价键在生理温度下就能不断地形成和破坏。结果，它们容易进行可逆的（暂时的）分子间相互作用，而这些作用是生理功能所必需的。对于核酸和蛋白质来说，它们发挥了全部种类的关键作用，确保 DNA 准确复制、RNA 转录、密码子-反密码子识别。虽然单个的非共价键的作用力较弱，但是许多非共价键联合作用能够对这些分子的结构稳定性（**构象**，conformation）

做出巨大的贡献，因此对描述一个大分子的形态至关重要［见图 1.7B 关于分子内氢键的形成如何提供一转移 RNA 分子的大部分形态的例子］。

1.2 DNA 的结构和复制

1.2.1 DNA 的结构是一个反向平行的双螺旋

如前所述，一个 DNA 分子和一个 RNA 分子的线性骨架由交替的糖残基和磷酸基组成。在每种情况下，连接一个单独的糖残基与相邻糖残基的键是 **3′，5′-磷酸二酯键**（3′，5′-phosphodiester bond）。这意味着一个磷酸基连接了一个糖的 3′碳原子与相邻的糖的 5′碳原子（图 1.4）。

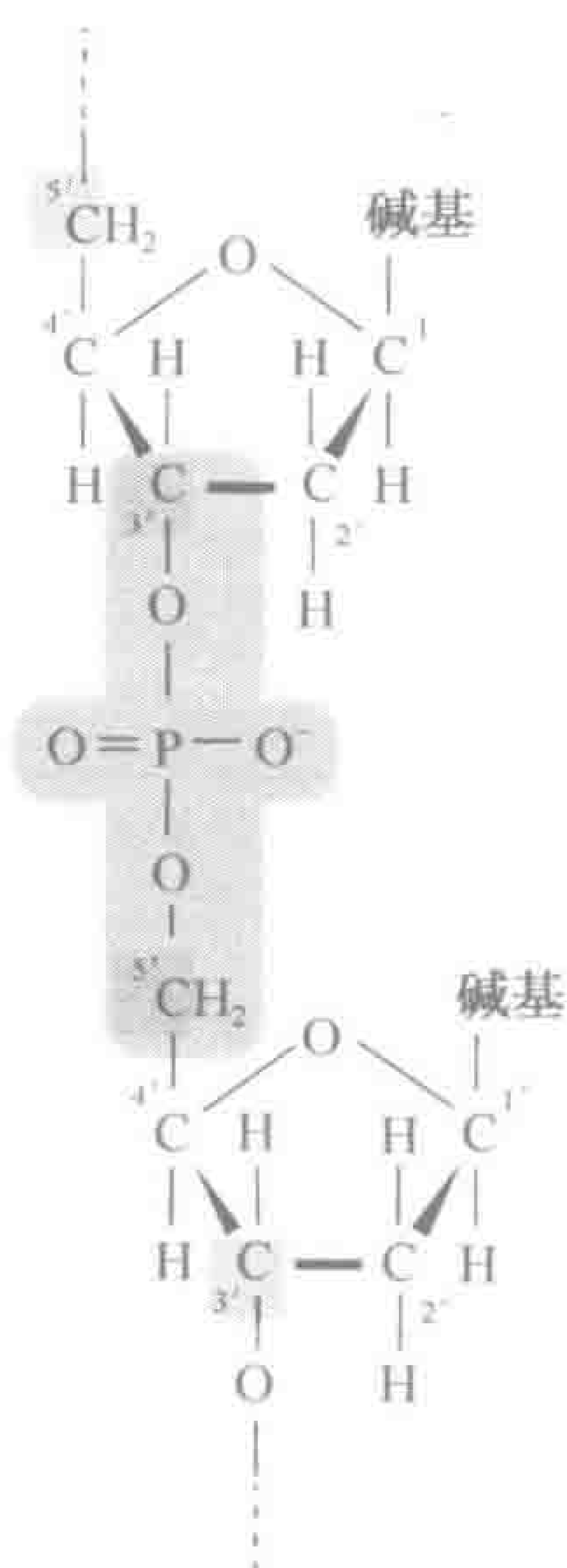


图 1.4　一个 3′-5′磷酸二酯键

一个细胞内的 RNA 分子通常以单个分子存在，而 DNA 的结构是一个**双螺旋**（double helix），在此双螺旋中，两个 DNA 分子通过弱氢键结合在一起形成一个 **DNA 双链**（DNA duplex）。依据 Watson-Crick 法则，氢键形成发生于 DNA 双链侧面相对的碱基，即**碱基对**（base pair，bp）之间。Watson-Crick 法则是指：A 特异地与 T 结合，C 特异地与 G 结合（图 1.5）。结果，来自不同细胞来源的 DNA 的碱基组分不是随机的：腺嘌呤与胸腺嘧啶数量相等，胞嘧啶与鸟嘌呤数目相等。因此，通过引用碱基的%GC（=%G+%C）组分，可以清楚地说明 DNA 的碱基组分。例如，如果一种细胞 DNA 描述为 42% GC，那么可以推断出碱基的组分为：G，21%；C，21%；A，29%；T，29%。

DNA 可以采用不同类型的螺旋结构。A-DNA 和 B-DNA 都是右手螺旋（右手螺旋是指当螺旋远离观察者时螺旋沿着顺时针方向盘旋）。每一圈，它们分别有 11bp 和 10bp。Z-DNA 是一个左手螺旋，每一圈有 12bp。在生理条件下，细菌和真核生物基因组的大多数 DNA 以 B-DNA 形式存在。在这里，每一条螺旋链具有 3.4nm 的**螺距**（pitch）（一圈螺旋所占有的距离）。因为磷酸二酯键连接了连续的糖残基的 3′和 5′碳原子，所以每条 DNA 链的一个末端，即所谓的 **5′端**（5′ end），含有一末端的糖残基，其中的 5′碳原子不与邻近的糖残基连接（图 1.6）。另一个末端由于类似的末端糖残基 3′碳原子缺少磷酸二酯键形成而定义为 **3′端**（3′ end）。据说一个 DNA 双链的两条链是反向平行的，因为它们总是以这样的方式结合，即按一条 DNA 链的 5′→3′方向与另一条链相反的方向结合（**复性**，anneal）（图 1.6）。

遗传信息由 DNA 链中碱基的线性序列（**一级结构**，primary structure）编码。因此，DNA 双链的两条 DNA 链具有**互补**（complementary）序列［或表现为**碱基互补性**（base complementarity）］。对于一条 DNA 链的碱基序列，若已知其互补链的 DNA 序列，则可以很容易地被推断出来。因此，通常通过仅仅写出一条链的碱基序列，按照

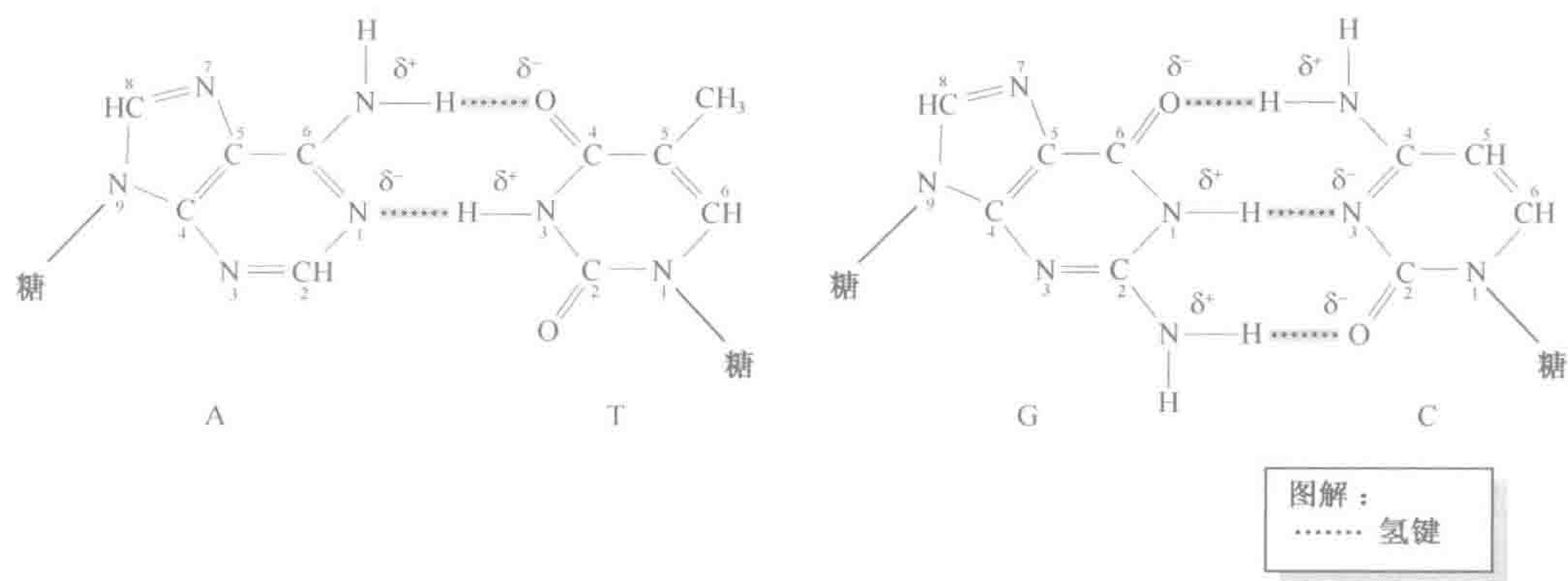


图1.5　A-T碱基对有两个连接的氢键；G-C碱基对有三个

氢原子的部分正电荷与氧原子和氮原子的部分负电荷分别用 δ^+ 和 δ^- 表示。

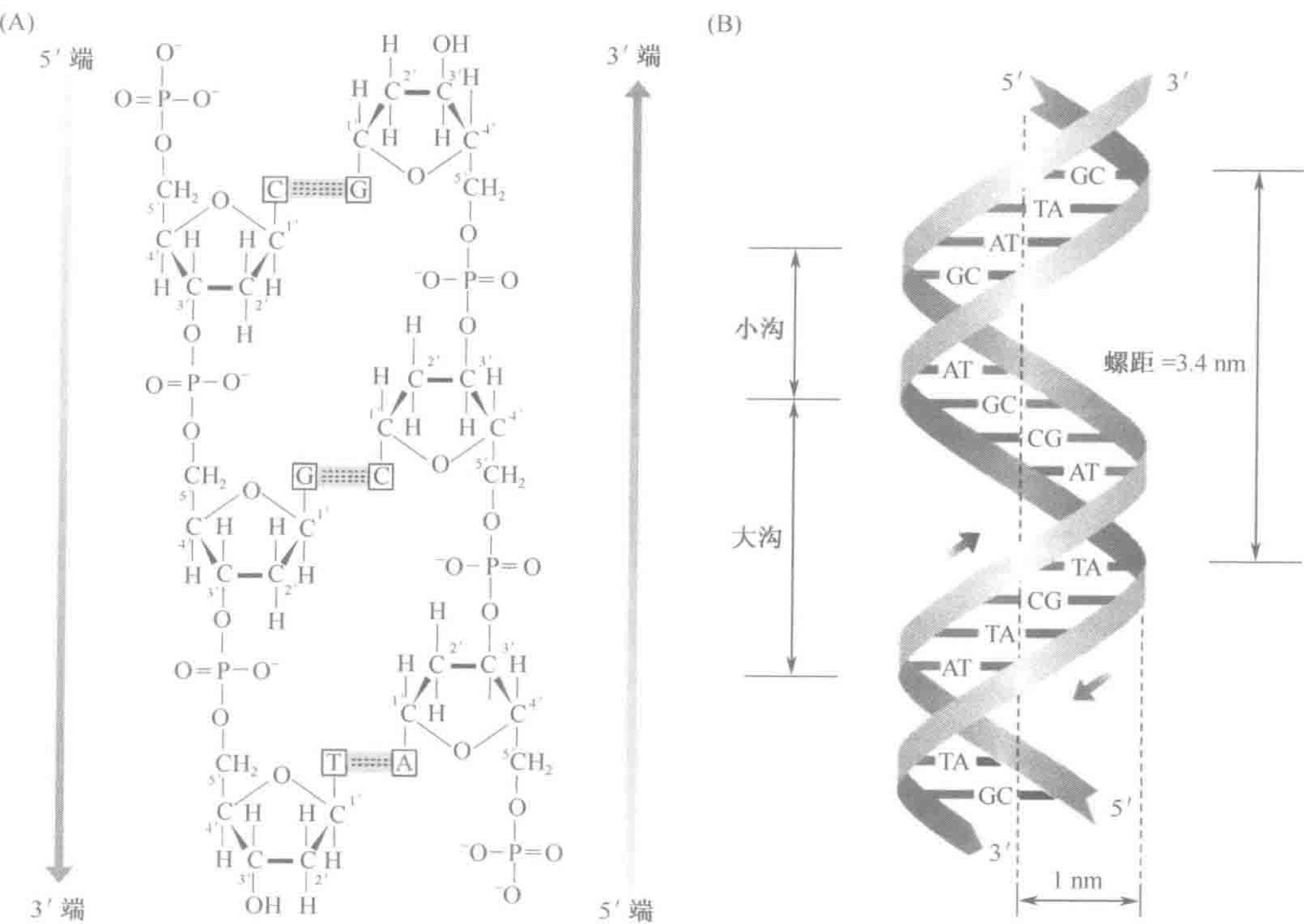


图1.6　DNA的结构是一个双链、反向平行的螺旋

(A) **两条DNA链的反向平行性质。**两条链是反向平行的，因为它们具有相反方向的3′碳原子与5′碳原子的结合。显示的结构是一个双链的核苷酸，其序列可表示为5′pCpGpT—OH3′（左侧DNA链）/5′pApCpG—OH3′（右侧DNA链）（此处p=磷酸二酯键，—OH=3′端最后的羟基）。这通常通过除去“p”和“OH”符号以及只给出一条链的序列而简略（例如序列可同样恰当地表示为5′CGT3′或5′ACG3′）。(B) **DNA的双螺旋结构。**两条链会彼此环绕形成一个相互缠绕的螺旋。每个螺旋的**螺距**（pitch）为一单个转角所占据的距离，在B-DNA结构（见正文）中容纳10个核苷酸。

5′→3′方向来描述一个 DNA 序列。此方向是 DNA 复制过程中新的 DNA 分子合成的方向，也是使用 DNA 作为模板合成 RNA 分子时转录的方向（见下文）。然而，当描述一条 DNA 链上包含两个相邻碱基（实际上是一个二核苷酸）的 DNA 区域的序列时，通常插入一个“p”来表示一个连接的磷酸二酯键，例如 CpG 表示在同一条 DNA 链上一个胞苷共价地连接到一个相邻的鸟苷上，而一个 CG 碱基对表示一条 DNA 链上的一个胞嘧啶与互补链上的一个鸟嘌呤形成氢键（图 1.6）。

分子间的氢键形成也允许 RNA-DNA 双链和双链 RNA 的形成，这些是基因表达重要的必要条件（框 1.1）。另外，氢键形成可发生于一单个 DNA 或 RNA 分子内的碱基之间。位置很近的互补反向重复序列倾向于形成**发夹**（hairpin）结构或袢，通过袢颈部碱基之间的氢键形成而稳定（图 1.7A；也见图 9.6 关于通过切割发夹 RNA 前体形成 microRNA 的例子）。这些结构限制，是一级结构所强加的限制之外的结构限制，有助于稳定分子的**二级结构**（secondary structure）。某些 RNA 分子，如转移 RNA（tRNA），显示出特别高程度的二级结构（图 1.7B）。

框 1.1　核酸与蛋白质中氢键形成重要性的例子

核酸中分子间氢键形成

这对于允许如下双链核酸形成很重要：

- **双链 DNA**。双螺旋的稳定性是通过 A-T 和 G-C 碱基对之间氢键形成来维持的（1.2.1 节和图 1.5）；
- **DNA-RNA 双链**。这些在 RNA 转录过程中自然形成，氢键形成支持下述类型的碱基配对：A-U，C-G，也有 A-T（RNA 链的 A 与 DNA 链的 T 之间的键形成）（节 1.3.3 和图 1.12）。
- **双链 RNA**。一些病毒基因组由 RNA-RNA 双链组成，但在所有细胞内的 RNA 加工及基因表达过程中还有暂时的 RNA-RNA 双链形成。RNA 剪接需要外显子-内含子界限的识别，随后在未剪接的 RNA 转录物与不同的小核 RNA 分子间形成氢键（节 1.4.1）。另外，密码子-反密码子识别涉及两个 RNA 分子（mRNA 和 tRNA）之间的氢键形成（节 1.5.1 和图 1.20）。RNA 分子间的氢键形成涉及 G-U 碱基配对，也涉及 A-U 与 G-C 碱基配对（表 1.6）；

核酸中分子内氢键形成。这对于在 DNA 和 RNA 分子中提供二级结构很重要，就像在 DNA 中发夹结构的形成以及复杂的 tRNA 臂一样（节 1.2.1 和图 1.7）。在后一种情况，注意碱基配对既涉及G-U碱基配对，也涉及 A-U 与 G-C 碱基配对。

蛋白质中分子内氢键形成。一些蛋白质二级结构的基本单位，如 α 螺旋、β 折叠与 β 转角，主要由链内氢键形成所限定（节 1.5.5 和图 1.24）。

注：对于 RNA-RNA 双链的氢键形成以及 RNA 分子内氢键形成来说，除了 A-U 和 C-G 碱基配对之外偶尔发现 G-U 碱基配对（图 1.7B）。这种形式的碱基配对不是特别地稳定，但是不会显著地破坏 RNA-RNA 螺旋。

1.2.2　DNA 复制是半保留的而 DNA 链的合成是半不连续的

在 DNA 合成（**DNA 复制**，DNA replication）的过程中，每条染色体的两条 DNA 链由解旋酶展开，每一条 DNA 链指导一互补 DNA 链的合成，形成两个子 DNA 双链，每一个都与亲代分子完全相同（图 1.8）。因为每一个子 DNA 双链包含一条亲代

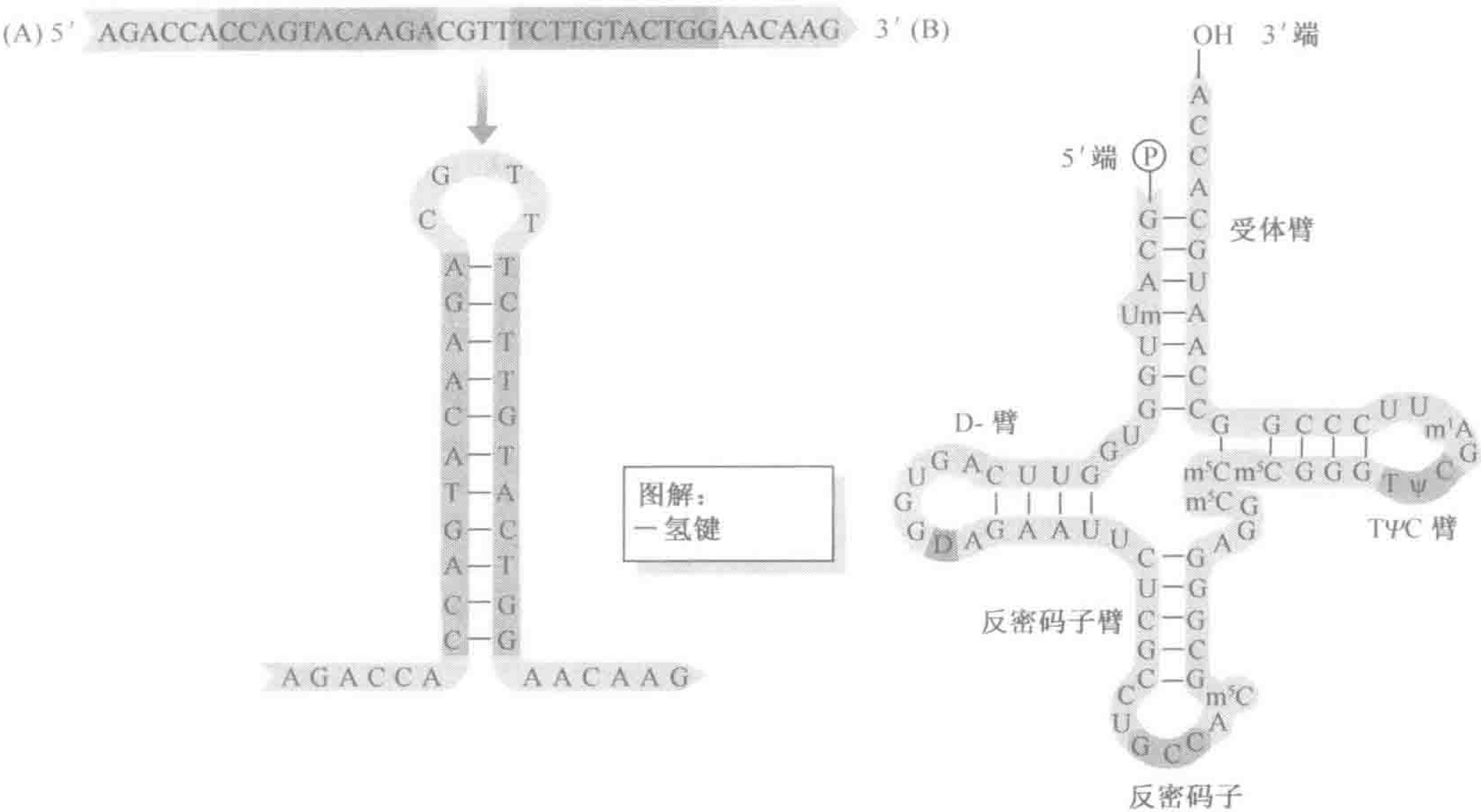


图 1.7　DNA 和 RNA 中分子内氢键形成

（A）**一单个 DNA 链内一个发夹环的形成**。上面 DNA 链内突出显示的序列为反向重复序列，能够通过氢键形成一发夹结构（见下面）。（B）**转移 RNA（tRNA）具有广泛的二级结构**。显示的例子为一个人类 $tRNA^{Glu}$ 基因。次要核苷有：D，5，6-二氢尿嘧啶核苷；Ψ，假尿嘧啶（5-核糖尿嘧啶）；m^5C，5-甲基胞苷；m^1A，1-甲基腺苷。三叶草结构由于广泛的分子内氢键形成而稳定，大多数通过 Waston-Crick G-C 和 A-U 碱基配对，也有偶然的 G-U 碱基配对。四个臂得到公认：**受体臂**（acceptor arm）为可附加一个氨基酸（在 3′端）的臂；**TΨC 臂**由此三核苷酸限定；**D 臂**因其含有二氢尿嘧啶核苷残基而得名；**反密码子臂**（anticodon arm）在环的中心部含有反密码子三核苷酸。tRNA 的二级结构实际上是不变的：总有七个碱基对位于受体臂的茎，五个碱基对位于 TΨC 臂，五个碱基对位于反密码子臂，以及三个或四个碱基对位于 D 臂。

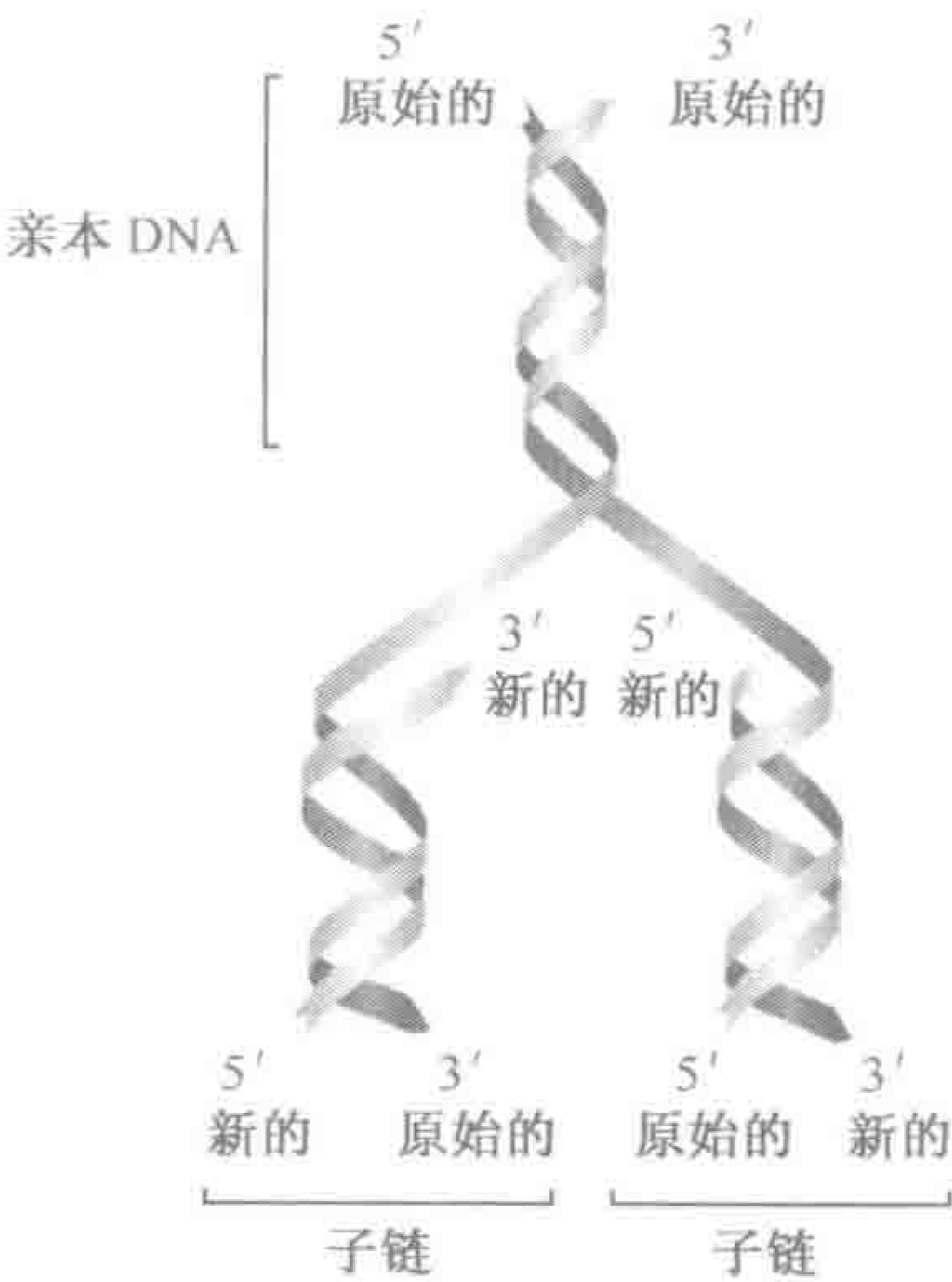


图 1.8　DNA 复制是半保留的

亲代 DNA 双链由两条互补反向平行的 DNA 链组成，这两条链解开，随后各自作为模板用于新的互补反向平行 DNA 链的合成。每个子 DNA 双链含有一条原始的亲代 DNA 链和一条新 DNA 链，形成一个结构上与亲代 DNA 双链相同的 DNA 双链。**注**：此图显示了 DNA 复制的结果而不是过程进行的方式（关于此内容见图 1.9）。

分子和一条新合成的 DNA 链，所以复制过程被描述为**半保留的**（semi-conservative）。DNA 聚合酶利用四种脱氧核苷三磷酸（dATP，dCTP，dGTP，dTTP）作为核苷酸前体，催化新 DNA 链的合成。

DNA 复制起始于特定的点，这些点被命名为**复制起点**（origin of replication）。从这些起点开始，DNA 复制的起始形成一个 Y 形**复制叉**（replication fork），亲代 DNA 双链在此分叉（分裂）进入两个子 DNA 双链。亲代 DNA 双链的两条链是反向平行的，但独自作为模板用于合成一条互补的反向平行的子链。结果，两条子链一定是沿相反方向伸展［即：一条子链——**前导链**（leading strand）——的链增长方向一定是 5′→3′；但另一条子链——**后随链**（lagging strand）——的链增长方向是 3′→5′；图 1.9］。

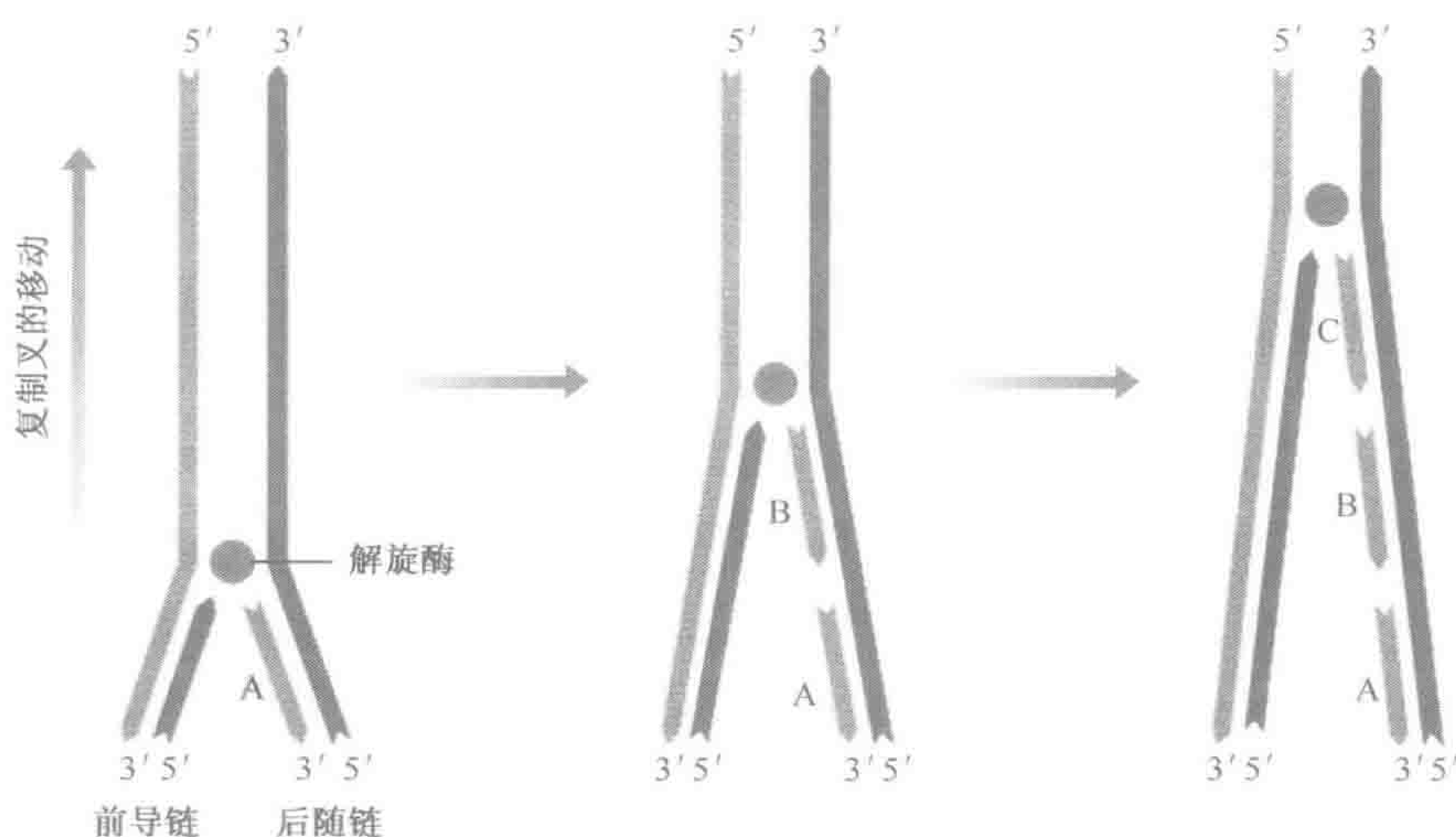


图 1.9　DNA 复制过程中链合成的不对称性

解旋酶解开 DNA 双链，使得各个链均被复制。当复制进行时，前导链合成的 5′→3′方向与复制叉移动的方向相同，所以合成是连续的。然而，后随链的合成与复制叉移动的方向相反。它需要以碎片（冈崎片段）形式合成，首先为 A，然后为 B，然后为 C，随后由 DNA 连接酶封闭并形成一连续的 DNA 链。**注**：这些片段的合成是利用一 RNA 引物开始的。在真核细胞中，前导链和后随链是分别通过 DNA 聚合酶 δ 和 α 来合成的（表 1.2）。

DNA 聚合酶催化的反应包括添加一个 dNMP 残基至正在增长的 DNA 链的游离 3′羟基（dNMP 残基由一个 dNTP 前体提供——两个末端的 β 和 γ 磷酸残基被切割而剩下的焦磷酸基团被丢掉）。这个必要条件给 DNA 复制过程引入了一个不对称性：只有前导链在分叉点有一个游离的 3′羟基。这将允许核苷酸的相继添加以及沿着复制叉移动方向的连续延伸。

后随链 5′→3′的合成方向与复制叉移动方向相反。因此，合成不得不分步进行，生成一渐进系列的小片段（**冈崎片段**，Okazaki fragment），一般为 100～1000 个核苷酸长。因为只有前导链是连续合成的，所以 DNA 链的合成被称做**半不连续的**（semi-discontinuous）。后随链的每一个片段都是沿着 5′→3′方向合成，这个方向与复制叉移动的方向相反。利用 DNA 连接酶，将接连合成的片段在其末端共价地连接到一起。结果，后随链沿复制叉移动的方向延长。

1.2.3 哺乳动物细胞的DNA复制装置是复杂的

就像核糖体一样，虽然有各种关键的不同蛋白质类型，从大肠杆菌到哺乳动物细胞，DNA复制的机制是高度保守的，（框1.2）。然而，无论是从不同DNA聚合酶的数目，还是从组成的蛋白质和蛋白质亚单位的数目，哺乳动物细胞中DNA复制的复杂性更大。在哺乳动物细胞里，大多数DNA聚合酶利用一单独的DNA链作为合成一条互补DNA链的模板，因此它是DNA介导的DNA聚合酶。与RNA聚合酶不同，DNA聚合酶绝对需要一碱基配对的引物链的3′羟基末端作为链延伸的底物。因此，需要一个预先合成的RNA引物（由引发酶合成）来提供DNA聚合酶开始合成DNA所需的游离3-OH基团。

框1.2 用于DNA复制机构的主要类型的蛋白质

拓扑异构酶（topoisomerase）通过在一单个DNA链上打开缺口（切断）起始DNA解旋过程。结果，以螺旋和超螺旋形式维持双螺旋的张力被释放。

一旦一个拓扑异构酶去除了超螺旋，那么**解旋酶**（helicase）则完成原始双链的展开。

DNA聚合酶（DNA polymerase）合成新的DNA链。DNA聚合酶是几个不同的蛋白质亚单位的复杂聚集体，通常具有DNA**即时校读**（proof-reading）和**核酸酶**（nuclease）活性，因此任何错误掺入的碱基都能被识别，并且含有这样错误的局部DNA链能被切除，随后修复。在细胞中，DNA一般通过利用DNA介导的DNA聚合酶从一个已存在的DNA链模板进行新的DNA合成而复制的，哺乳细胞中存在多种这样的DNA聚合酶。在更专有的情况下，DNA可在细胞中利用RNA介导的DNA聚合酶（也称为反转录酶）从RNA模板中合成，就像利用**端粒酶**（telomerase）的反转录酶活性合成一线性染色体末端一样。

引发酶（primase）。DNA难以从一裸露的单链模板开始从头合成。反之，需要一个引物，聚合酶能够在其3′羟基基团附加一个dNTP。引发酶在单链DNA附加一个小的RNA引物，作为DNA聚合酶合成开始的一个替代的3′羟基。这个RNA引物最后由一个核糖核酸酶去除，缺口利用一DNA聚合酶填充，随后由DNA连接酶封闭。

连接酶（ligase）催化既定的，未附着而邻近的3′羟基与5′磷酸间形成一磷酸二酯键。

单链结合蛋白（single-stranded binding protein）对于维持复制叉的稳定性很重要。单链DNA是非常易变的或者不稳定的，因此当它保持单链时这些蛋白质与其结合，防止其降解。

在哺乳动物细胞内有20种多种不同类型的DNA聚合酶（Friedberg *et al.*，2000；2002），这些聚合酶可方便地分成三大类（表1.2）：

表1.2 哺乳动物DNA聚合酶的主要类型

(A) 高保真性（经典的）DNA介导的DNA聚合酶

	DNA聚合酶				
	α	β	γ	δ	ε
位置	细胞核	细胞核	线粒体	细胞核	细胞核
普遍的DNA复制	后随链的合成与引发	—	mtDNA复制	前导链的合成	—

续表

	DNA 聚合酶				
3′→5′外切核酸酶[a]	无	无	有	有	有
DNA 修复功能	—	通过碱基切除	mtDNA 修复	通过核苷酸和碱基切除	通过核苷酸和碱基切除

[a]用作一个即时校读活性

(B) 低保真性（有错误倾向的）DNA 介导的 DNA 聚合酶

DNA 聚合酶 ζ（zeta）	在突变的 B 细胞和 T 细胞中表达，参与高突变?
DNA 聚合酶 η（eta）	在高突变过程中使 A 和 T 核苷酸产生突变?
DNA 聚合酶 ι（iota）	非常低的复制保真性；认为参与高突变
DNA 聚合酶 μ（mu）	在 B 细胞和 T 细胞中高表达；认为参与高突变

(C) RNA 介导的 DNA 聚合酶（反转录酶）

端粒酶反转录酶（Tert）	复制线性染色体末端的 DNA
LINE-1/内源性反转录病毒反转录酶	偶尔将 mRNA 及其他 RNA 转变为能够整合至基因组其他地方的 cDNA

- **经典的（高保真性）DNA 介导的 DNA 聚合酶。**在这些酶中，两种涉及标准的染色体 DNA 复制，特异性用于从前导链或后随链合成 DNA（分别为 DNA 聚合酶 δ 和 α），两种致力于 DNA 修复（β 和 ε），还有一种致力于复制和修复线粒体 DNA（γ）；
- **有错误倾向的 DNA 介导的 DNA 聚合酶**。已经鉴定了各种不同的、具有非常低的 DNA 复制保真性的 DNA 聚合酶［例如 DNA 聚合酶 ι（iota）的错误率是 DNA 聚合酶 ε 的 20 000 倍］。已知一些酶在免疫系统细胞中极高地表达，这些观察加上它们 DNA 复制的低保真性提示其参与 B 和 T 淋巴细胞的高突变（节 10.6）；
- **RNA 介导的 DNA 聚合酶**。一些 DNA 聚合酶使用一个 RNA 模板来合成 DNA，因此这些酶被描述为**反转录酶**（reverse transcriptase）（节 1.2.4 和节 1.3.1）。它们包括存在于端粒酶（端粒酶负责线性染色体末端的复制）的活性（节 2.2.5），以及由一些高度重复的 DNA 和内源性反转录病毒类型编码的反转录酶。

1.2.4 病毒基因组往往由 RNA 复制而不是由 DNA 复制来维持

DNA 是现今所有细胞的遗传物质，而且我们习惯于把**基因组**（genome）看作是一个有机体或细胞的遗传 DNA 分子集合的专有名词。尽管作为细胞遗传物质的 DNA 无处不在，但是许多不同类型的现存病毒具有 RNA 基因组。RNA 分子可进行自我复制，但是 RNA 核糖残基上的 2′羟基基团使糖-磷酸键在化学上相当不稳定。在 DNA 中，脱氧核糖残基在 2′位置上只带有一个氢原子，因此 DNA 比 RNA 更适于作为一个稳定的遗传信息的载体。RNA 复制也是易于出错的，正常的 RNA 复制错误率大约是在 DNA 复制过程中出现的错误率的 10 000 倍。

病毒已形成了许多不同的策略来感染和破坏细胞，并且它们的基因组表现出惊人的多样性（表 1.3；图 1.10）。由于它们有很高的突变负荷，所以病毒 RNA 基因组一般很小但具有高速突变率的优点。与 DNA 病毒不同（DNA 病毒一般在核内复制），RNA 病毒通常在细胞质中复制。在这一普遍规则的某些重要的例外中有一组称为**反转录病毒**（retrovir-use）的 RNA 病毒。反转录病毒是一种罕见的 RNA 病毒，它们在核内复制，并且使用**反转录酶**（reverse transcriptase）经由一种 DNA 中间体进行复制。反转录酶是一 RNA 介导

的 DNA 聚合酶，与 RNA 聚合酶一样，具有相对高的复制错误率。RNA 转变为互补 DNA（cDNA）后，反转录病毒 cDNA 就整合至宿主的染色体 DNA 中（节 21.5.3）。

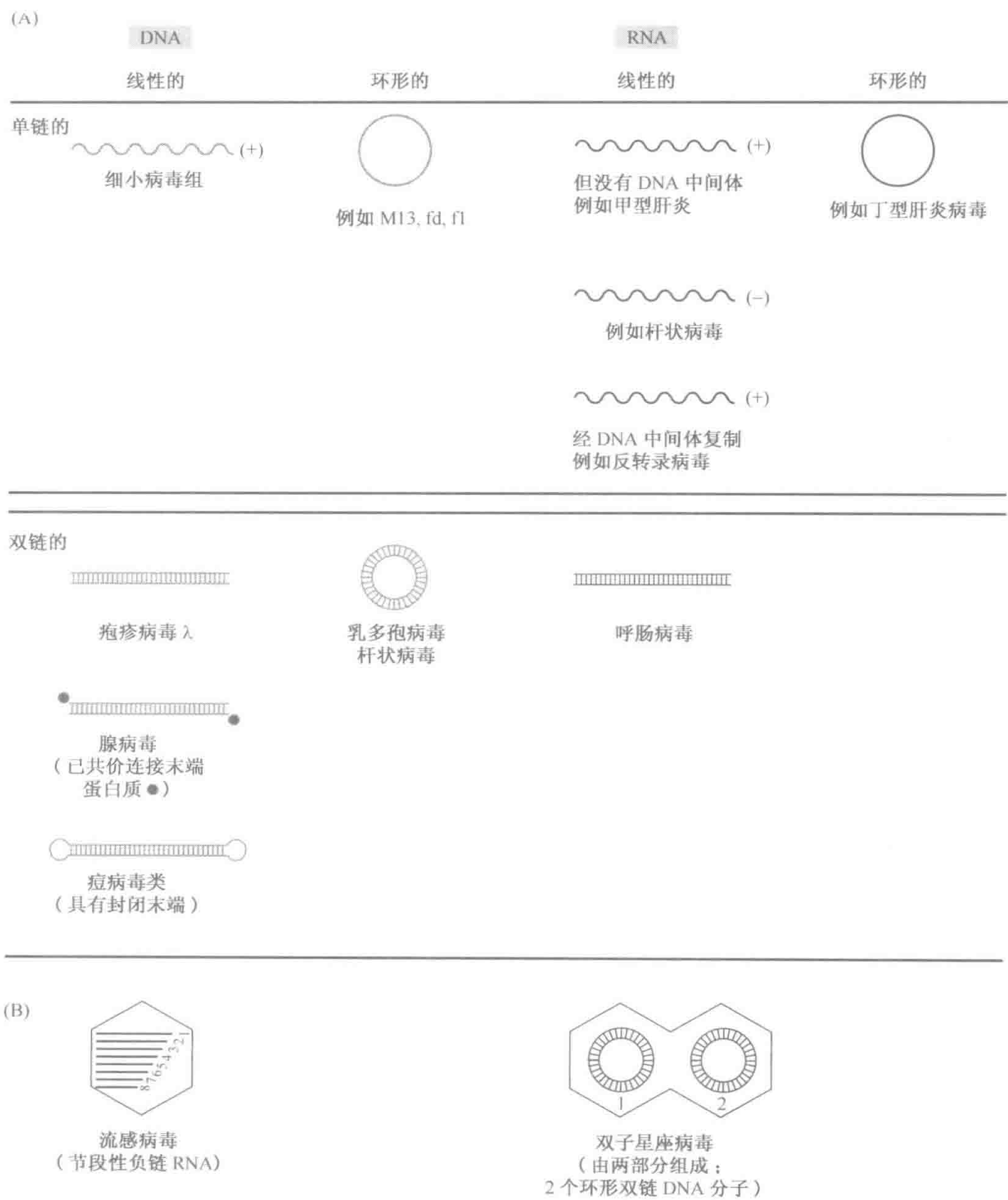


图 1.10　非常多样化的病毒基因组

（A）**病毒基因组的成链性与拓扑学**。对于单链病毒基因组来说，用于产生蛋白质产物的 RNA 可以具有与基因组相同的有义链，因此称为一个正链基因组（+），或者具有与基因组相反的链（反义链），称为一个负链基因组（−）。一些单链（+）RNA 病毒经过一个 DNA 中间体［**反转录病毒**，（retrovirus）］，而一些双链 DNA 病毒诸如乙型肝炎经过一个复制 RNA 形式。（B）**节段性和多部分的病毒基因组**。节段性基因组具有各种不同的、指定单顺反子 mRNA 的核酸分子。就像对于流感病毒来说，它具有 8 个不同的负单链 RNA 分子。多部分基因组是节段性基因组的一个亚类，每一个不同的分子包装成一个独立的病毒颗粒。

表 1.3 不同类型的基因组（病毒基因组结构的例子见图 1.10）

	DNA		RNA	
	双链（ds）	单链（ss）	双链（ds）	单链（ss）
单个环状分子	许多细菌和古细菌；线粒体；叶绿体；一些病毒	一些病毒	—	非常少的病毒
单个线性分子	非常少的细菌，例如 Borrella；一些病毒	一些病毒	少数病毒	一些病毒
多个线性分子	真核生物细胞核；一些节段性双链 DNA 病毒	一些节段性单链 DNA 病毒	一些节段性双链 RNA 病毒	一些节段性单链 RNA 病毒
多个环状分子	一些细菌；一些有两个部分和有三个部分的病毒	—	—	—
线性和环状的混合	非常少的细菌，例如根癌农杆菌	—	—	—

1.3 RNA 转录和基因表达

1.3.1 细胞中遗传信息流几乎是专有的一个途径：DNA→RNA→蛋白质

所有细胞中遗传信息的表达是一个非常主要的单向系统：DNA 特化 RNA 的合成，然后 RNA 特化多肽的合成（随后形成蛋白质）。由于其通用性，所以从 DNA→RNA→多肽（蛋白质）的遗传信息流称为分子生物学的**中心法则**（central dogma）。在所有细胞有机体中，两个连续的步骤是必不可少的：

1. **转录**（transcription）。这是利用一个 DNA 介导的 RNA 聚合酶，发生于真核细胞的细胞质中。在有限的程度上，发生于线粒体和叶绿体中，除了细胞核之外这是唯一具有遗传能力的其他细胞器（图 1.11）；

2. **翻译**（translation）。这发生在核糖体中。核糖体是大的 RNA-蛋白质复合物，常见于细胞质以及线粒体和叶绿体中。指定多肽的 RNA 分子称为**信使 RNA**（messenger RNA）。

遗传信息的表达遵循**共线性原理**（colinearity principle）：DNA 中核苷酸的线性序列每次以三核苷酸组（**碱基三联体**，base triplet）形式被解码形成 RNA 中核苷酸的线性序列，该序列又依次以三核苷酸组（**密码子**，codon）形式被解码形成多肽产物中氨基酸的线性序列。

最近，逐渐清楚的是真核细胞包括哺乳动物细胞含有编码细胞反转录酶的非病毒染色体 DNA 序列，如哺乳动物 LINE-1 重复 DNA 家族的成员。因为已知一些非病毒 RNA 序列可用作细胞 DNA 合成的模板，所以细胞中遗传信息单向流动的原理不再绝对健全。

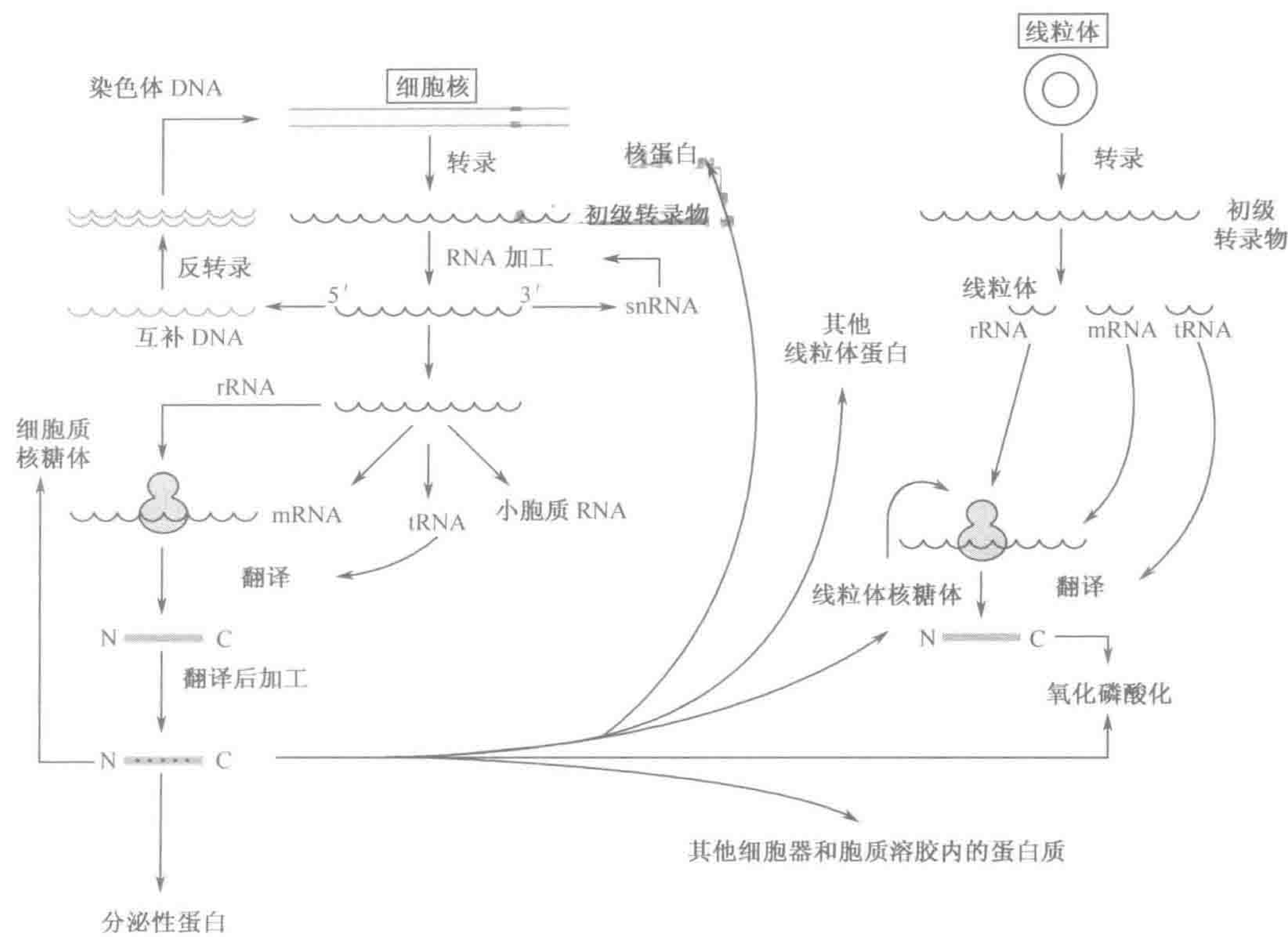


图 1.11　一个动物细胞内基因表达

注：（Ⅰ）非常偶然地，一小部分核 RNA 分子能够自然地通过病毒编码的或细胞的反转录酶转换为 cDNA，随后在不同位置整合至染色体 DNA；（Ⅱ）线粒体合成它自己的 rRNA、tRNA 和几个涉及氧化磷酸化系统的蛋白质。然而，线粒体核糖体蛋白、线粒体氧化磷酸化系统中大多数蛋白质以及其他线粒体蛋白在输入线粒体之前是由核基因编码并在细胞质核糖体上翻译。

1.3.2　在复杂的有机体内只有一小部分 DNA 表达并形成一个蛋白质或 RNA 产物

细胞中所有的 DNA 只有一小部分被转录。根据它们的需要，不同细胞转录不同的 DNA 节段（**转录单位**，transcription unit），这些转录单位是不连续的单位，沿 DNA 序列不规则地分布。RNA 聚合酶以转录单位为模板合成初始的相等大小的 RNA 分子（**初级转录物**，primary transcript），然后初级转录物被修饰，产生成熟的表达产物。然而，在任一细胞中绝大多数细胞 DNA 是从来不转录的。而且，只有一部分由转录产生的 RNA 翻译成多肽。这是因为：

- 一些转录单位特化非编码 RNA：此 RNA 产物不编码多肽如 mRNA，但具有不同的功能。除了完全确定的核糖体 RNA（rRNA）和转移 RNA（tRNA）之外，我们现在知道各种具有不同功能的非编码 RNA（节 9.2）；
- 特化为 mRNA 的转录单位的初级转录物经历 RNA 加工。结果，很多最初的 RNA 序列被丢弃，形成一更小的 mRNA（节 1.4.1）；
- 成熟 mRNA 只有中心部分被翻译；mRNA 每个末端可变长度部分始终是非翻译的

（节 1.5.1）。

在动物细胞中，DNA 存在于细胞核和线粒体中。但是，线粒体只含有很小部分的总细胞 DNA 以及非常有限数目的基因（9.1.2）；一个细胞的绝大多数 DNA 位于细胞核的染色体中。

复杂真核细胞基因组中**编码 DNA**（coding DNA）部分相当小。这部分是由于基因内有许多非编码性质序列的结果。另一个原因是相当大部分的复杂真核细胞基因组含有没有功能的或不转录成 RNA 的重复序列。前者包括功能基因的缺陷性拷贝（假基因和基因片段）以及高度重复非编码 DNA。

1.3.3　在转录过程中一些 DNA 节段（基因）的遗传信息特化 RNA

RNA 合成是利用 RNA 聚合酶，以 DNA 为模板，以 ATP、CTP、GTP 和 UTP 为 RNA 前体而完成的。RNA 是作为一条单链合成的，合成方向为 5′→3′。通过将适当的核苷一磷酸残基（AMP、CMP、GMP 或 UMP）添加至不断增长的 RNA 链 3′端的游离羟基上，发生链延长。这些核苷酸是从适当的核苷三磷酸（rNTP）前体分离一个焦磷酸残基（PPi）得到的。这意味着最远的 5′端核苷酸（**起始核苷酸**，initiator nucleotide）与链内所有其他的核苷酸不同，携带一个 5′三磷酸基团。

通常地，两条 DNA 链只有一条用作 RNA 合成的模板。在转录过程中，双链 DNA 展开，而用作 RNA 合成模板的 DNA 链与增长的 RNA 链形成一个暂时的**双链 RNA-**

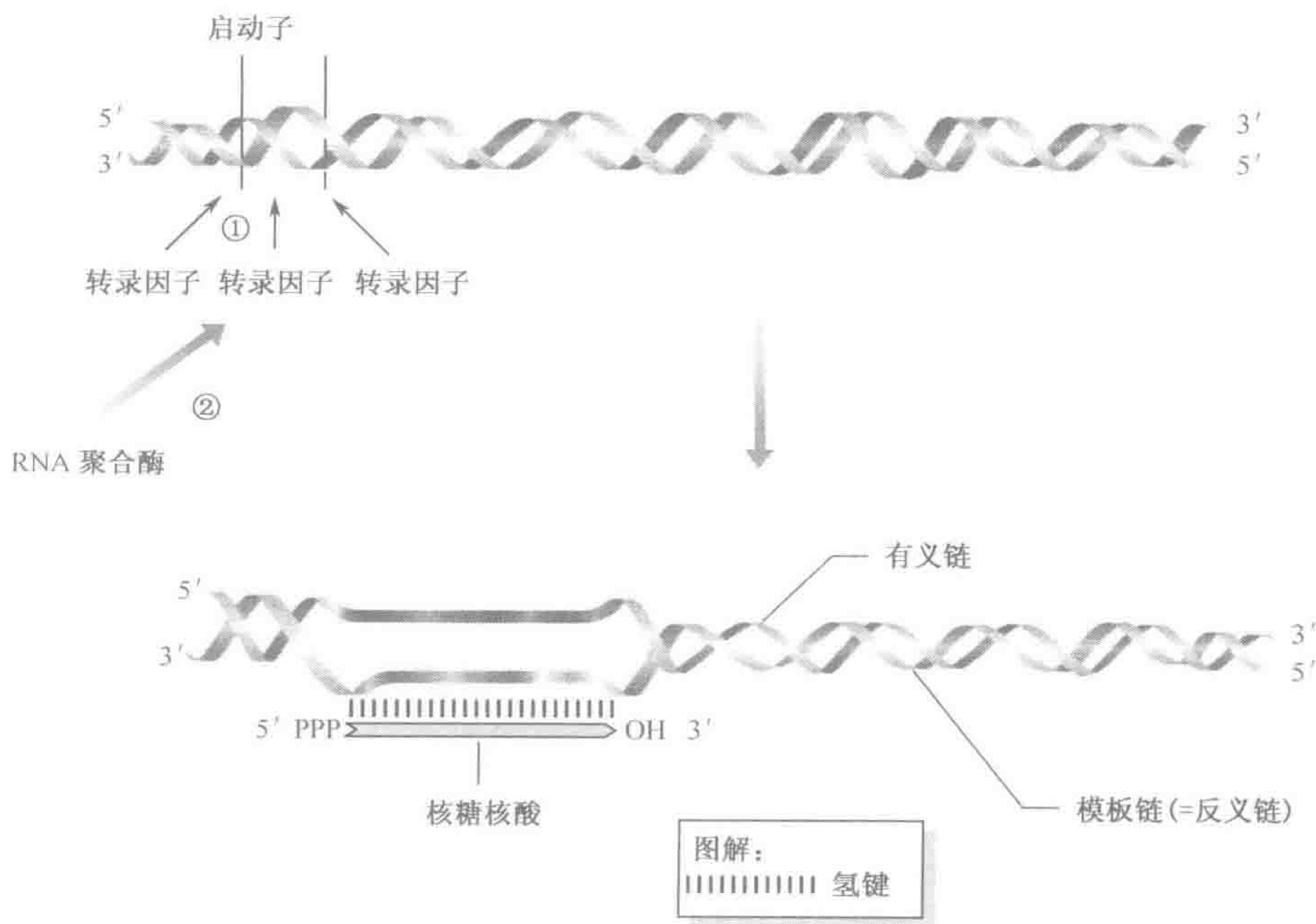


图 1.12　RNA 作为一条单链转录，与一个基因的一条链（模板链）在碱基序列上互补
需要各种转录因子结合于一个基因邻近的启动子序列①，为的是随后定位和指导转录基因的 RNA 聚合酶②。链合成开始于一个核苷三磷酸而链延长则通过连续地添加由 rNTP 提供的核苷一磷酸至 3′羟基发生。这意味着 5′端将有一个三磷酸基团（随后可经历修饰，如加帽；节 1.4.2），而 3′端将有一个自由的羟基。
注：RNA 序列一般与基因的有义链相同（除了 U 代替 T），而在序列上与模板链互补。

DNA杂种（RNA-DNA hybrid）。由于RNA转录物与这个**模板链**（template strand）互补，所以转录物具有与相对的、双螺旋的非模板链一样的方向和碱基序列（除了U代替T）。由于这个原因，非模板链通常称为**有义链**（sense strand）而模板链通常称为**反义链**（antisense strand）（图1.12）。记录基因序列时习惯于只显示有义链的DNA序列。与一个基因序列有关的序列方向通常是指有义链的方向。例如，一个基因的5′端是指位于有义链5′端的序列，而**上游**（upstream）或**下游序列**（downstream sequence）是指参照有义链，分别位于基因5′或3′端旁侧的序列。

在真核细胞中，三种不同的RNA聚合酶分子是合成不同类型RNA所必需的（表1.4）。绝大多数细胞基因编码多肽并由RNA聚合酶Ⅱ转录。然而，越来越多的重要性给予了编码RNA作为其成熟产物的基因：现在已知功能性RNA分子具有各种作用，而催化功能已被归为其中（节9.2.3）。

表1.4　三类真核生物RNA聚合酶

类型	转录的基因	注释
Ⅰ	28S rRNA；18S rRNA；5.8S rRNA	定位于细胞核。一单个初级转录物（45S rRNA）可被切割产生列举的三种rRNA类型。
Ⅱ	编码多肽的所有基因；大多数snRNA基因	唯有聚合酶Ⅱ转录物经历加帽和多腺苷酸化
Ⅲ	5S rRNA；tRNA基因；U6 snRNA；7SL RNA；7SK RNA；7SM RNA；SiRNA	由RNA聚合酶Ⅲ转录的一些基因（例如5S rRNA，tRNA，7SL RNA）的启动子位于基因的内部（图1.13），而其他基因的启动子（例如7SK RNA）则位于上游。

1.3.4　顺式作用调节元件和反式作用转录因子是真核基因表达所必需的

真核生物RNA聚合酶不能独自起始转录。反而，位于一个基因直接邻近处的短序列元件的组合可用作**转录因子**（transcription factor）与DNA结合的识别信号，以便指导和激活聚合酶。一组主要的这样的短序列元件通常簇集于一个基因编码序列的上游，在那里它们集合构成一个**启动子**（promoter）。许多通用转录因子与启动子区域结合之后，RNA聚合酶与转录因子复合物结合并被激活，从一个单独的位置起始RNA合成。据说转录因子是**反式作用**（*trans*-acting）的，因为它们是由遥远定位的基因所合成的，并需要迁移到它们发挥作用的位置。相反，据说启动子是**顺式作用**（*cis*-acting）的，因为它们的作用仅限于它们存在的DNA双链（表1.5）。主要的顺式作用元件的功能性分组包括：

表1.5　由普遍存在的转录因子所识别的顺式作用元件的例子

顺式元件	与其一致的或是其变异体的DNA序列	相关的反式作用因子	注释
GC框	GGGCGG	Spl	Spl因子是普遍存在的
TATA框	TATAAA	TFⅡD	TFⅡA与TFⅡD-TATA框复合物结合，使其稳定

续表

顺式元件	与其一致的或是其变异体的DNA序列	相关的反式作用因子	注释
CAAT 框	CCAAT	许多,例如 C/EBP;CTF/NF1	反式作用因子大家族
TRE(TPA 反应元件)	GTGAGT(A/C)A	AP-1 家族,例如 JUN/FOS	反式作用因子大家族
CRE(cAMP 反应元件)	GTGACGT(A/C)A(A/G)	CREB/ATF 家族,例如 ATF-1	基因对 cAMP 应答而激活

▶ **启动子** (promoter)。常见的顺式作用元件包括：

- TATA 框，通常为 TATAAA 或一个变异体，常见于转录起始点上游大约 25bp（—25bp）处（图 1.13）。它一般见于由 RNA 聚合酶Ⅱ活跃转录的基因，但在有限的意义上说，是在细胞周期的特定时期（例如组蛋白）或者特定细胞类型（例如 β 珠蛋白）。TATA 元件的突变不妨碍转录的起始，但确实引起转录起始点从正常位置的移位。
- GC 框，一致性序列 GGGCGG 的一个变异体，见于各种基因，许多缺乏 TATA 框，如持家基因（节 1.3.5）。尽管 GC 框序列是非对称性的，但是它看起来好像在任何一个方向都发挥作用（图 1.13）。
- CAAT 框，常位于—80 位置。它通常是启动子效率最强的决定因子。与 GC 框一样，它看起来好像能够在任何一个方向都发挥作用（图 1.13）。

除了上述元件之外，已知更多特定的识别元件被组织限制性转录因子所识别（表 10.3）。

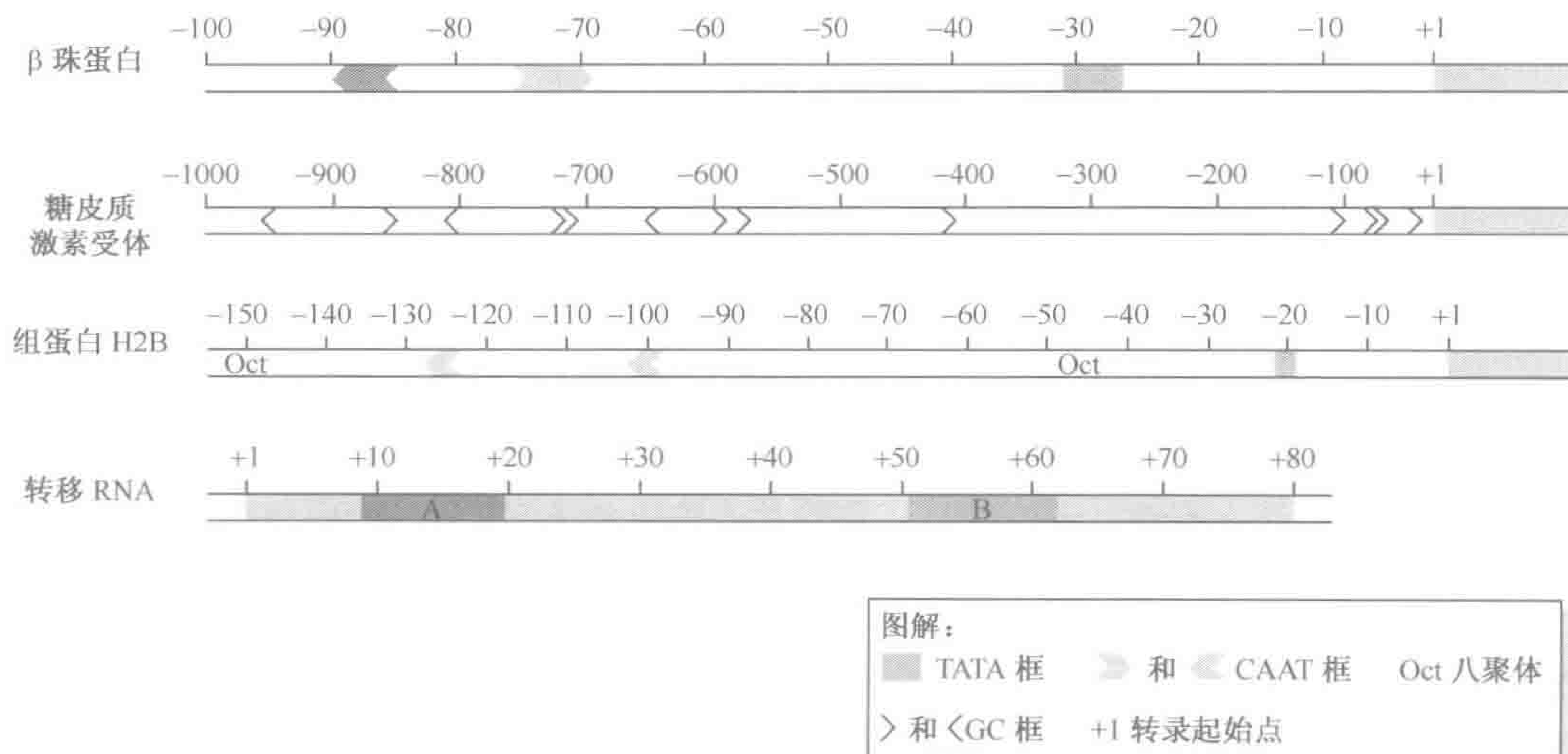


图 1.13 真核细胞启动子由距转录起始点相对恒定距离处的保守短序列的集合组成

GC 和 CAAT 框元件可选择的方向用人字形标识的方向指示；>表示正常方向；<表示相反方向。糖皮质激素受体基因与众不同的是拥有 13 个上游 GC 框（10 个位于正常方向；3 个位于相反方向）。tRNA 基因通过 RNA 聚合酶Ⅲ进行转录，并具有一个内部的由元件 A（根据标准的 tRNA 核苷酸编号系统，通常位于编号为+8～+19 的核苷酸内）和元件 B（通常位于+52～+62 核苷酸）两部分组成的启动子。特异性转录因子与这些元件结合并随后指导 RNA 聚合酶Ⅲ在+1 处开始转录。

▶ **增强子**（enhancer）由顺式作用短序列元件组构成，可增强特定真核基因的转录活性。然而，与启动子（相对于转录起始点的位置比较恒定）（图1.13）不同的是，增强子位于一个距离转录起始点可变的、通常相当远的位置，并且它们的功能不依赖于它们的方向。他们看起来好像与基因调节蛋白结合，随后启动子和增强子之间的DNA形成环，使得增强子结合的蛋白质与启动子结合的蛋白质，或者与RNA聚合酶相互作用。

▶ **沉默子**（silencer）与增强子具有相似性质，但却抑制特定基因转录活性的调节元件。

1.3.5 组织特异性基因表达涉及特定基因的选择性激活

一特定类型的真核细胞如肌细胞，其DNA含量实际上与淋巴细胞、肝细胞或来自同一有机体的任何其他类型的有核细胞的DNA含量一致。使不同细胞类型存在差异的是在任何一个细胞中只有一部分基因显著表达，而表达的基因目录在不同细胞类型间变化。在一些细胞中，尤其是脑细胞，表达大量的基因；而在一些其他细胞类型中，大部分基因是转录失活的。

显然，表达的基因是确定细胞功能的基因。一些功能是普遍的细胞功能，由所谓的**持家基因**（housekeeping gene）限定（例如编码组蛋白，核糖体蛋白等的基因）。其他功能可主要限制于特殊的组织或细胞类型（**组织特异性基因表达**，tissue-specific gene expression）。然而，值得注意的是，即使对于那些在表达上显示出相当大的组织特异性的基因来说，一些基因转录物可在所有细胞类型中表现极低的水平。

一个细胞中有转录活性和无转录活性DNA区域的区别反映在相关染色质的结构上：

▶ **转录失活的染色质**普遍采用一高度浓缩的构象，并通常与细胞周期S期中经历晚期复制的基因组区域相关。它通过紧密结合与组蛋白H1分子连接；

▶ **转录活性的染色质**采用一更开放的构象，并通常在S期早期复制。其标志是与组蛋白H1分子相对较弱的结合以及四种核小体组蛋白（即组蛋白H2A、H2B、H3和H4；节10.2.1）的广泛乙酰化。

另外，脊椎动物基因启动子区域的普遍特征是缺乏甲基化半胱氨酸（见下文）。转录因子可替代核小体，因此具有转录活性的染色质的开放构象可通过试验来辨别，因为它也为核酸酶提供了通路：在非常低的浓度下，脱氧核糖核酸酶Ⅰ（DNase Ⅰ）将消化长的无核小体DNA区域。尽管调节区域可含有几个序列特异性结合蛋白，但是开放性染色质结构的标志是**DNase Ⅰ 高敏感位点**（DNase Ⅰ hypersensitive site）的存在(节10.5.2)。

1.4 RNA加工

大多数真核基因的RNA转录物经历一系列加工反应。这通常包括除去不需要的内部节段以及重新连接剩余的节段（**RNA剪接**，RNA splicing）。另外，对于RNA聚合酶Ⅱ转录物来说，一专门的核苷酸连接（**7甲基鸟苷三磷酸**，7-methylguanosine

triphosphate）被添加至初级转录物的 5′端（**加帽**，capping），而腺苷酸（AMP）残基被相继地添加至 mRNA 的 3′端，形成一个 poly（A）尾（**多腺苷酸化**，polyadenylation）。

1.4.1 RNA 剪接从初级转录物中除去非必需的 RNA 序列

一个转录单位的线性序列支配一个相应的线性表达产物的合成，该表达产物要么是一个多肽，要么是一个成熟的非编码 RNA。然而，对于大多数脊椎动物基因来说，只有一小部分基因序列被解译，形成终产物。反而，在绝大多数编码多肽的基因和一些特定非编码 RNA 的基因中，遗传信息存在于间插序列所分隔的节段（**外显子**，exon）内，该间插序列不参与合成终产物的遗传信息（**内含子**，intron）。

起始的转录事件涉及与基因全长互补的 RNA 序列的产生，该 RNA 序列即所谓的**初级转录物**（primary transcript）。对于含有多个外显子的基因来说，初级转录物含有与基因内、外显子和内含子都互补的序列。然而，此后 RNA 转录物经历 **RNA 剪接**（RNA splicing），它是一系列加工反应，借此内含子 RNA 片段被剪掉并抛弃而外显子片段被端对端地连接（剪接的）从而产生一较短的 RNA 产物（图 1.14）。RNA 剪接需要识别位于外显子/内含子交界处（**剪接点**，splice junction）的核苷酸序列。在绝大多数情况下，内含子以 GT 开始（在 RNA 水平是 GU）并以 AG 结束（GT-AG 法则，图 1.15）。尽管保守的 GT（GU）和 AG 二核苷酸对于剪接非常重要，但它们独自不足以标志一个内含子的存在。引证序列的比较揭示出紧密邻近 GT 和 AG 二核苷酸的序列也显示出相当大程度的保守（图 1.15）。第三个已知的、在剪接中具有重要功能的保守内

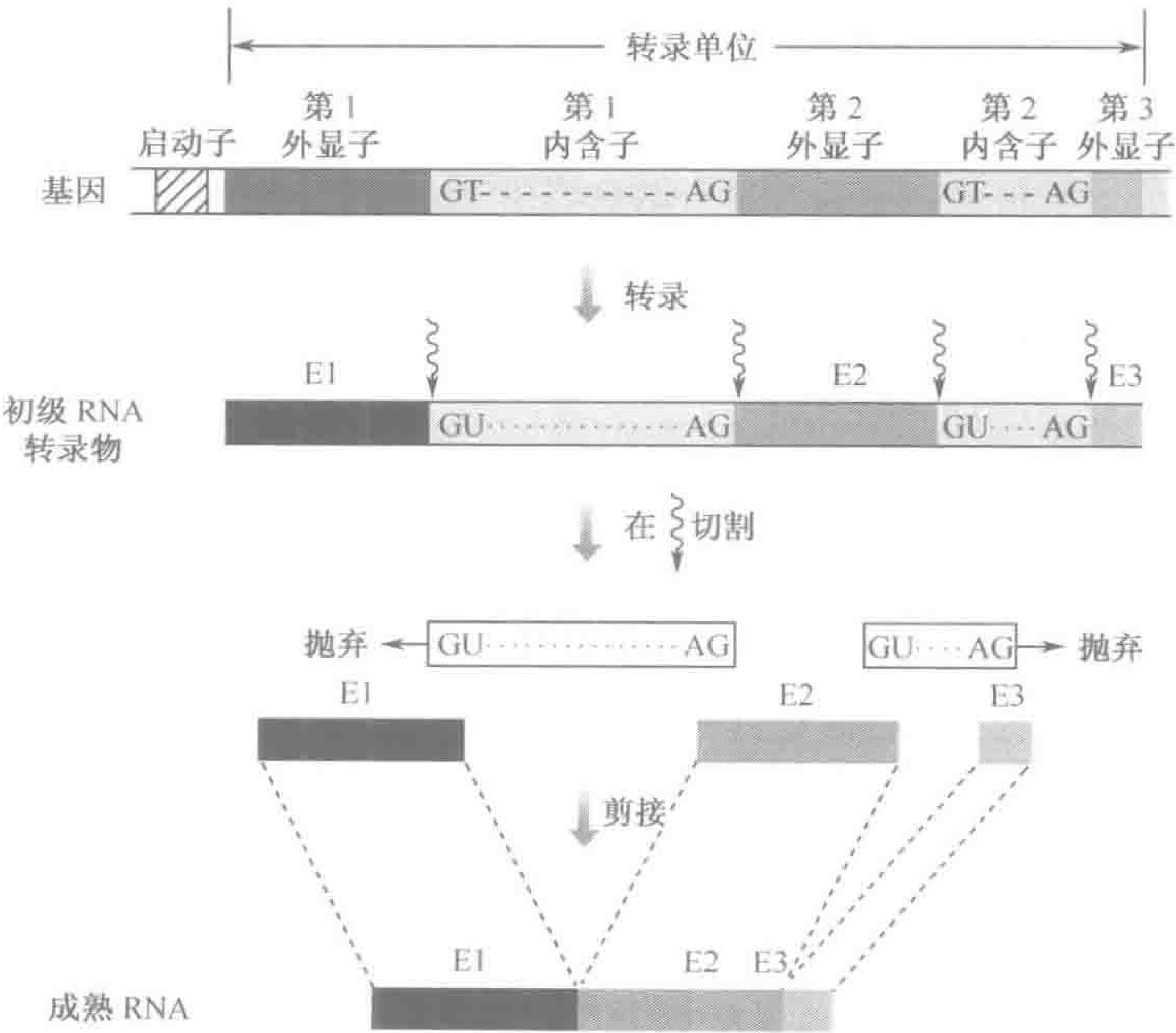


图 1.14　RNA 剪接涉及核内水解性切割、内含子 RNA 片段的去除以及外显子 RNA 片段的剪接

含子序列是所谓的**分支位点**（branch site），它通常位于非常邻近内含子末端的位置，在末端AG二核苷酸之前最多有40个核苷酸（图1.15）。另外，一些其他外显子和内含子序列对剪接具有正作用（**剪接增强子**，splice enhancer序列）或副作用（**剪接沉默子**，splice silencer序列），这些序列的突变是可以致病的（节11.4.3）。

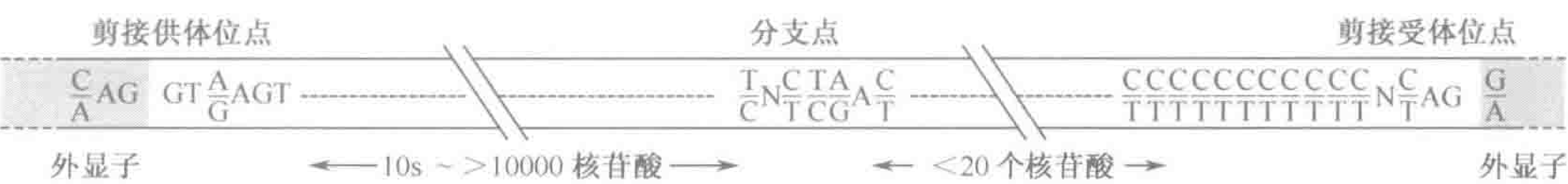


图1.15　复杂真核生物内含子中剪接供体、剪接受体及分支点在DNA水平的一致性序列

突出显示核苷酸是几乎不变的（注：也存在罕见的AT-AC内含子，在此内含子中，保守的剪接供体二核苷酸GT由AT代替而保守的剪接受体二核苷酸AG由AC代替；见正文）。其他核苷酸代表了在此特定位置的大多数核苷酸，或者两个核苷酸之间如邻近内含子3′端**多聚嘧啶束**（polypyrimidine tract）的C与T之间的核苷酸。Lim和Burge（2001）列举了各个物种（包括人类、果蝇、秀丽新小杆线虫、酿酒酵母）剪接供体，分支位点和剪接受体的优先基序。已知一些其他的外显子和内含子序列调节剪接，包括剪接增强子序列和剪接沉默子序列（Berget *et al.*，1995）。外显子剪接增强子序列的一致序列也参见Fairbrother等。(2002)。

剪接机制涉及下列序列：

- 在5′剪接点切割。
- 在分支点不变的A通过剪接供体位点末端G核苷酸亲核攻击，形成一个**套索**（lariat）状结构。
- 在3′剪接点切割，导致作为套索的内含子RNA的释放，以及外显子RNA片段的剪接（图1.16A）。

上述反应是由一个大的RNA-蛋白质复合物——**剪接体**（splicesome）介导的，而剪接体由五种**核内小RNA**（small nuclear RNA，snRNA）和50多种蛋白质组成（Staley and Guthrie，1998；Hastings and Krainer，2001）。每个snRNA分子附加于特定的蛋白质形成**核小核糖核蛋白颗粒**（snRNP particle），剪接反应的特异性是通过RNA转录物与snRNA分子间RNA-RNA碱基配对而建立的。有两类剪接体：

- **主要（GU-AG）剪接体**处理经典的GT-AG内含子。对于绝大多数编码多肽的基因的内含子来说，五种snRNA是：U1、U2、U4、U5和U6 snRNA。U1 snRNA 5′端具有一个与剪接供体一致序列（GUAAGUA）进行碱基配对的UACUUAC序列。U1 snRNP结合之后，U2 snRNA通过一个类似的碱基配对反应识别分支点，接着U1 snRNP和U2 snRNP之间的相互作用将两个剪接点紧密连接在一起。此后，一含有U4、U5和U6 snRNA的多snRNP颗粒与U1-U2 snRNP复合物结合在一起（图1.16B）。
- **次要（AU-AC）剪接体**处理一种罕见类型的内含子（称为AT-AC内含子）。因为保守的5′GT和3′AG二核苷酸分别被AT和AC二核苷酸替代。在这种情况下，U11和U12 snRNA代替了U1和U2（Tarn and Steitz，1997）。

剪接体被设想为以连续性方式发挥作用：一旦一个5′剪接点被识别，那么它就扫描RNA序列直到它遇到下一个3′剪接点（会被刚好位于它前面的分支位点一致序列标记为一个靶标）。然而，内含子序列被除去而其旁侧外显子序列被剪接的顺序并不是由

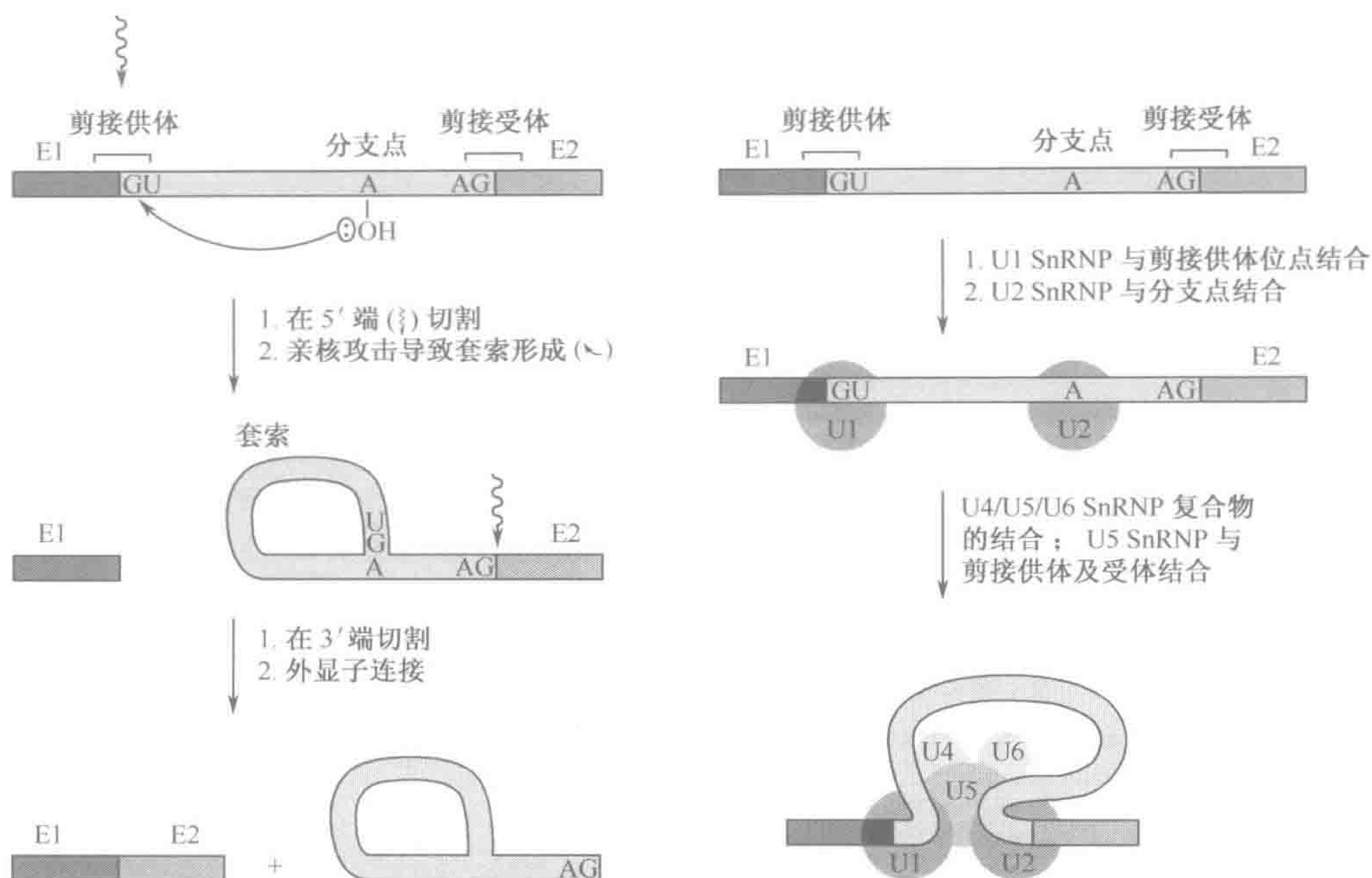


图 1.16 RNA 剪接（GU-AG 内含子）机制

（A）**机制**。剪接供体、剪接受体与分支位点的一致性序列见图 1.15。亲核攻击涉及分支点保守的 A 处附加的 2′羟基与内含子起始处保守的 GU 中的 G，并产生一个连接这两个核苷酸的新共价键，形成一个分支状结构（套索）。（B）**snRNP 作用**。核小核糖核蛋白颗粒（snRNP）是剪接体的一部分。U1 snRNA 具有一个与剪接供体互补的末端序列并通过 RNA-RNA 碱基配对与其结合。U1 snRNP 结合之后，U2 snRNA 通过一个相似的碱基配对反应识别分支位点。剪接供体与剪接受体接合处的相互作用通过随后与一个预先形成的多 snRNP 颗粒结合而稳定，这个多 snRNP 颗粒含有 U4、U5、U6 snRNA，其中 U5 snRNP 能够同时与剪接供体和剪接受体结合。

RNA 转录物中它们的线性顺序所支配的，反之认为 RNA 的构象影响 5′剪接位点的可及性。

1.4.2 大多数 RNA 聚合酶Ⅱ转录物的 5′和 3′端添加特殊的核苷酸

除了 RNA 剪接之外，RNA 聚合酶Ⅱ转录物经历两个额外的 RNA 加工事件：

加帽

这发生于转录后不久。对于将要加工形成 mRNA 的初级转录物来说，一个甲基化的核苷——7 甲基鸟苷（m^7G）——通过一特有的 5′-5′磷酸二酯键连接至 RNA 转录物的第一个 5′核苷酸。因为这个键有效地桥接了 m^7G 残基 5′碳原子和第一个核苷酸的 5′碳原子，所以 5′末端被封闭或加帽（图 1.17）。snRNA 基因的转录物也被加帽，但它们的帽可经历额外的修饰。曾设想帽具有几种可能的功能：

- 保护转录物免受 5′→3′外切核酸酶的攻击（脱帽的 mRNA 迅速降解）；
- 促进从细胞核向细胞质的转运；
- 促进 RNA 剪接；

▶ 在胞质核糖体 40S 亚基与 mRNA 的结合中发挥重要作用（见下文）。

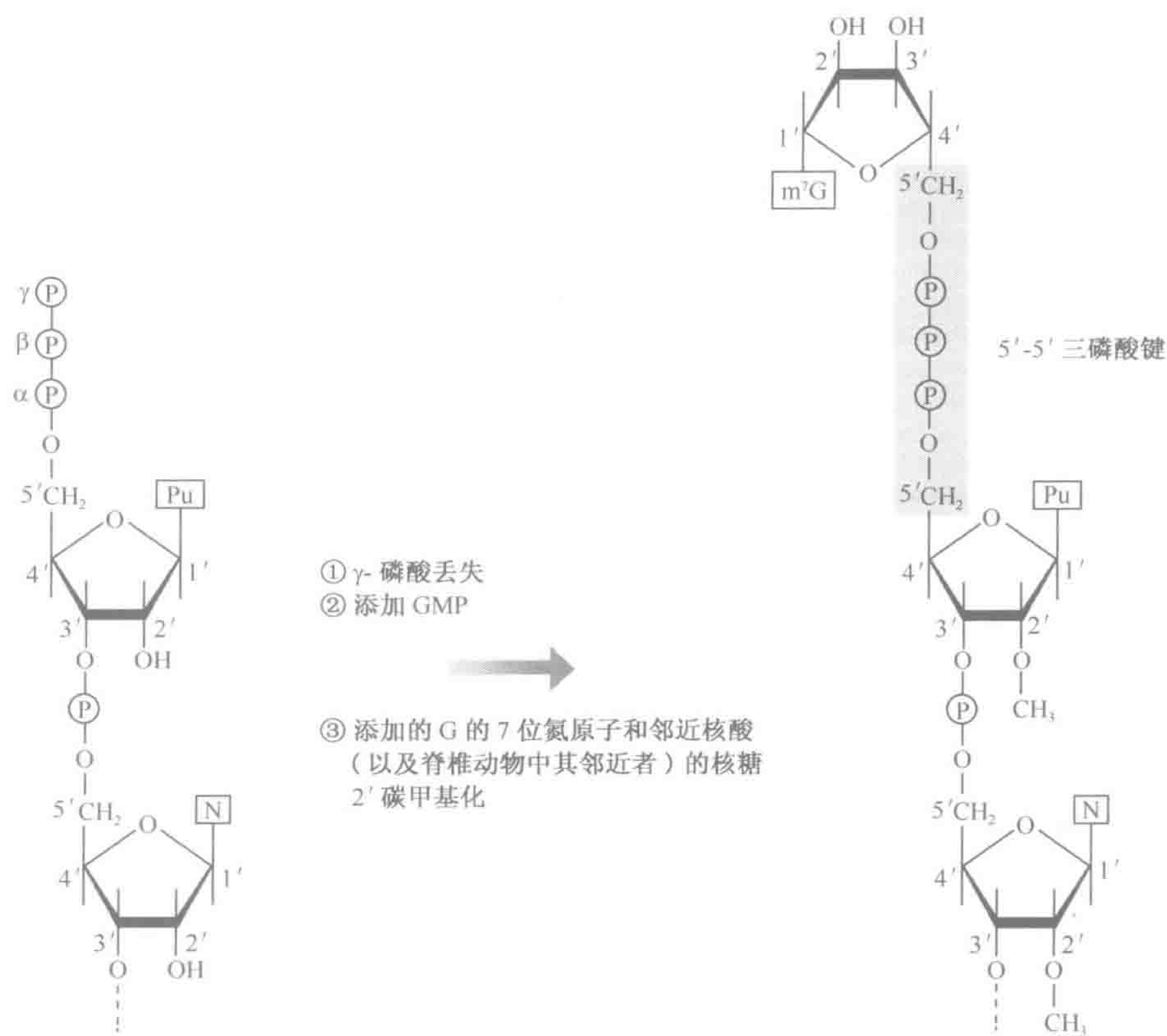


图 1.17　真核生物 mRNA 分子 5′端受一个特殊的核苷酸保护（加帽）

除去末端 5′核苷酸原始的 γ 磷酸后，一个 GTP 前体提供了一个新的 GMP 残基，与末端 5′核苷酸形成一个特化的 5′-5′三磷酸键。随后的反应导致末端 G 的 7 位氮原子的甲基化，在脊椎动物中为两个邻近核苷酸中每一个核苷酸的核糖的 2′碳原子甲基化。N，任一核苷酸；Pu，嘌呤。

多腺苷酸化

已知通过 RNA 聚合酶Ⅰ和Ⅲ进行的转录在酶识别一特定的转录终止位点之后就停止了。然而，对于通过 RNA 聚合酶Ⅱ进行的转录来说，很难识别可能的转录终止位点，因为 mRNA 分子的 3′端是由转录后切割反应所决定的。AAUAAA 序列（或者有时是 AUUAAA 变异体）是主要的**多腺苷酸化信号序列**（polyadenylation signal sequence），标志绝大多数聚合酶Ⅱ转录物的 3′切割（组蛋白基因和 snRNA 基因的转录物是明显的例外）。切割发生在位于 AAUAAA 元件下游 15～30 个核苷酸的特定位置（图 1.18）。

初级转录物可持续经过切割点几百或几千个核苷酸，直到在几个稍后位置之一发生终止。一旦 AAUAAA 元件下游发生切割，那么随后在哺乳动物细胞内通过 poly(A)聚合酶就可以添加大约 200 个**腺苷酸**（adenylate）（即 AMP）残基，形成一个 **poly(A)尾**（poly(A)tail）。曾设想该尾具有几种可能的功能：

▶ 促进 mRNA 分子转运至细胞质；

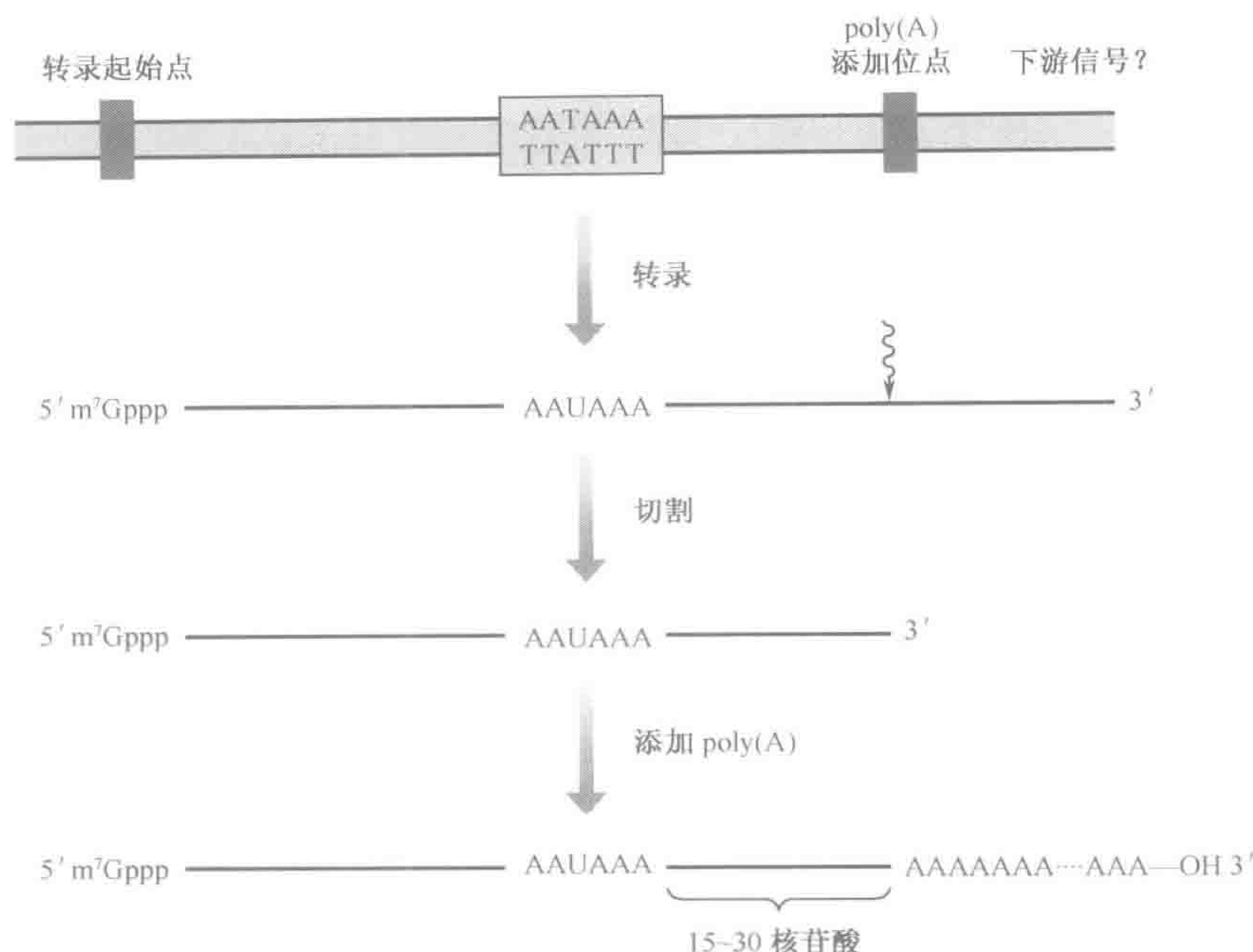


图 1.18 大多数真核生物 mRNA 分子 3′端是多腺苷酸化的

RNA 聚合酶Ⅱ转录物的转录末端以转录的 RNA 中一个 3′端切割为信号。在大多数 mRNA 种类中，这通过一个上游 AAUAAA 信号与仍旧未鉴定的下游信号一起完成。切割一般发生于 AAUAAA 元件下游大约 15～30 个核苷酸，随后通过 poly（A）聚合酶添加 AMP 残基形成一个 poly（A）尾。组蛋白 mRNA 经历一个不同的 3′端切割反应（见正文）。

- 至少在细胞质中稳定一些 mRNA 分子[poly(A)尾变短与 mRNA 降解有关，但是有些 mRNA 分子(例如肌动蛋白 mRNA)具有很短的或没有 poly(A)时仍旧很稳定]；
- 通过增强核糖体结构对 mRNA 的识别而促进翻译。

对于组蛋白基因来说，其独特之处在于产生未多腺苷酸化的 mRNA，而转录的终止也涉及初级转录物的 3′端切割。这个反应依赖于 RNA 转录物的二级结构，包括一个保守的上游发夹结构和一个短的下游序列，该序列可与 U7 snRNA 5′端的短序列碱基配对。

1.5 翻译、翻译后加工及蛋白质结构

1.5.1 翻译过程中 mRNA 在核糖体上解码以特化多肽的合成

转录后加工之后，核 DNA 基因转录的 mRNA 迁移至细胞质。在这里，它与核糖体及其他组分一起指导特定多肽的合成。线粒体也具有核糖体以及有限的蛋白质合成能力（节 9.1.2）。

只有一典型的真核生物 mRNA 分子的中心节段被翻译，以指导一个多肽的合成。旁侧序列，5′和 3′**非翻译区**（untranslated region，UTR）从 5′和 3′端外显子转录，与 5′帽和 3′poly(A)尾一样，它们有助于 mRNA 分子在核糖体（中心节段发生翻译）上的

结合与稳定（图 1.19）。

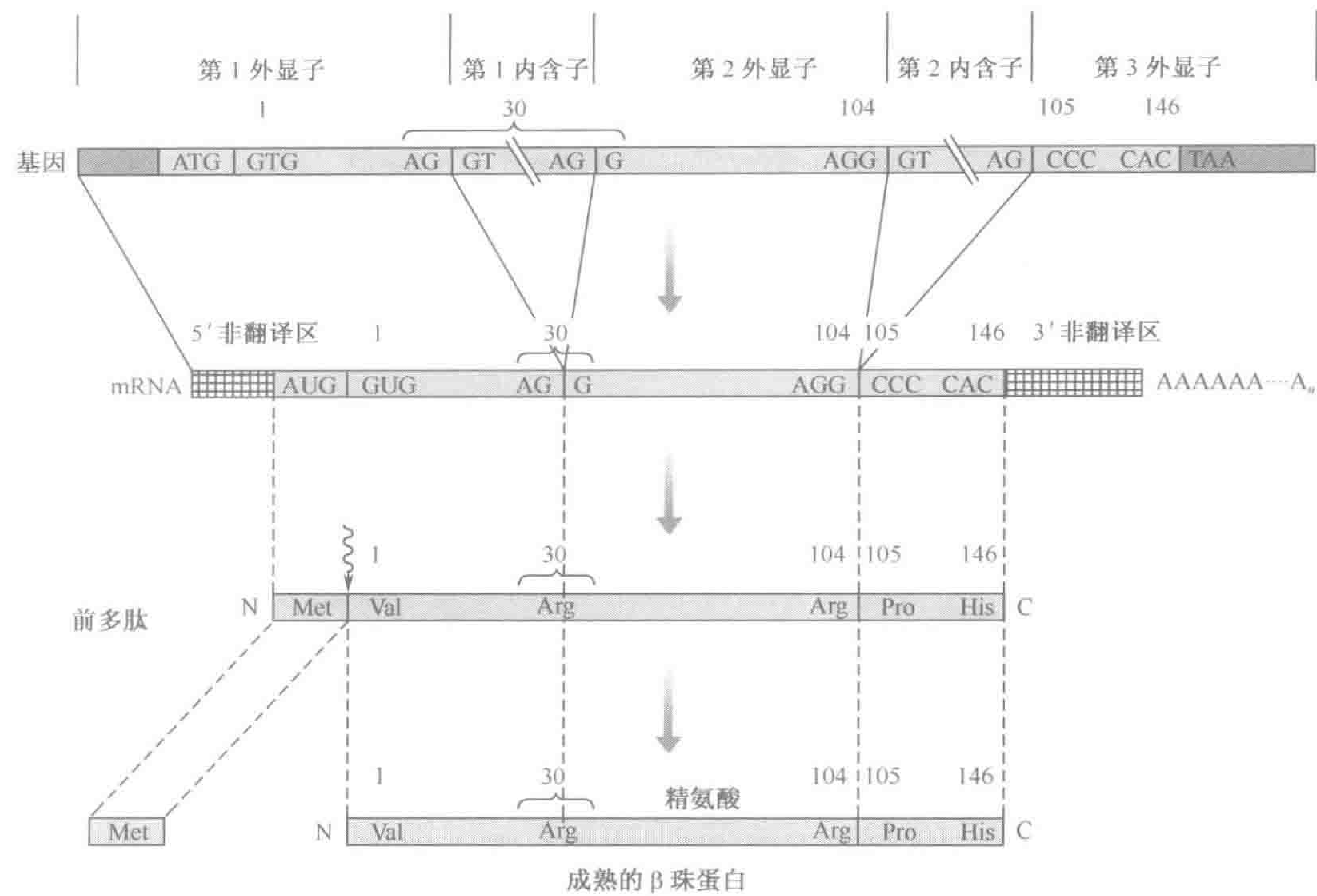


图 1.19 人类 β 珠蛋白基因的表达

第一外显子和第三外显子都在它们的末端含有非编码序列（▦），这些非编码序列可以转录并存在于 β 珠蛋白 mRNA 的 5′和 3′端，但并不翻译及指导多肽合成。然而，这些 5′和 3′非翻译区被认为在确保翻译的高效性中很重要（见正文）。终止密码子 UAA 是 3′非翻译区最初的三个核苷酸。**值得注意的是**最初的翻译产物有 147 个氨基酸，但是 N 端蛋氨酸通过翻译后加工被除去，产生成熟的 β 珠蛋白多肽。特定 Arg30 的密码子的前两个碱基由第一外显子编码而第三个碱基由第二个外显子编码（即第一内含子分离了密码子的第二个与第三个碱基，是一个 2 位相内含子，框 12.2；第二个内含子分离了密码子 104 与密码子 105，是一个 0 位相内含子的例子；一个 1 位相内含子的例子，框 12.2 与图 1.23）。

核糖体（ribosome）是大的 RNA-蛋白质复合物，由两个亚基组成。在真核细胞中，胞质核糖体有一个大的 60S 亚基和一个较小的 40S 亚基（S 值是大分子结构在超速离心时沉降快慢的一个衡量标准，受分子质量和形状的控制）。60S 亚基含有三种 rRNA 分子，28S rRNA、5.8S rRNA 和 5S rRNA，以及大约 50 种核糖体蛋白。40S 亚基含有一单个 18S rRNA 和 30 多种核糖体蛋白。核糖体为多肽合成提供了一个结构框架，其中 RNA 组分主要负责核糖体的催化功能。蛋白质组分被认为是增强 rRNA 分子功能的，而它们中令人吃惊的数目对核糖体的主要功能看起来并不重要。

一个新的多肽从其构成的氨基酸所进行的组装受**三联体遗传密码**（triplet genetic code）的控制。为特化各个氨基酸，线性 mRNA 序列上相继的三核苷酸组（**密码子**，condon）被连续地解码。解码过程由 tRNA 分子集合介导，每个 tRNA 分子已通过一特定的氨基酰 tRNA 合成酶共价结合了一个特定的氨基酸（在 tRNA 游离的 3′羟基；图 1.20）。

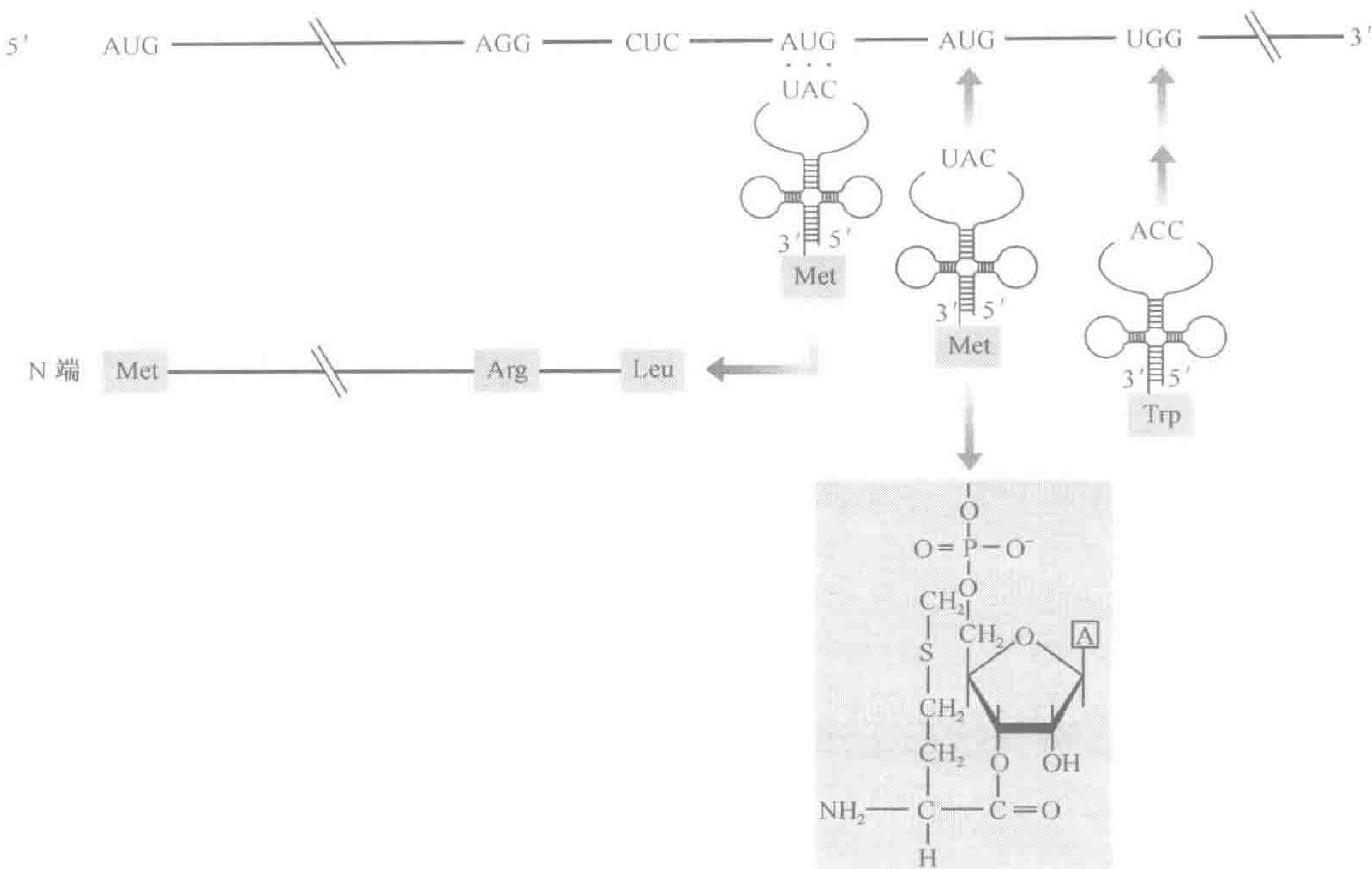


图 1.20　遗传密码通过密码子-反密码子识别而译解

mRNA 序列中核苷酸的序列从一翻译起始点（一般以序列 AUG 标示）开始翻译，并沿 5′→3′方向继续进行，直到到达那个可读框的终止密码子。mRNA 中每个密码子由一个 tRNA 分子的互补反密码子序列识别，该 tRNA 分子 3′端的腺苷共价结合某一特定氨基酸（见插入）。

不同的 tRNA 分子结合不同的氨基酸。每个 tRNA 具有一特定的三核苷酸序列，称为**反密码子**（anticodon），它位于 tRNA 分子一个臂中心非常重要的位置上（图 1.7B）。这个位置提供了解译遗传密码所必需的特异性：对于一个将插入不断增长的多肽链的氨基酸来说，mRNA 分子有关的密码子必须通过碱基配对被适当 tRNA 分子的适宜互补反密码子所识别（图 1.20）。

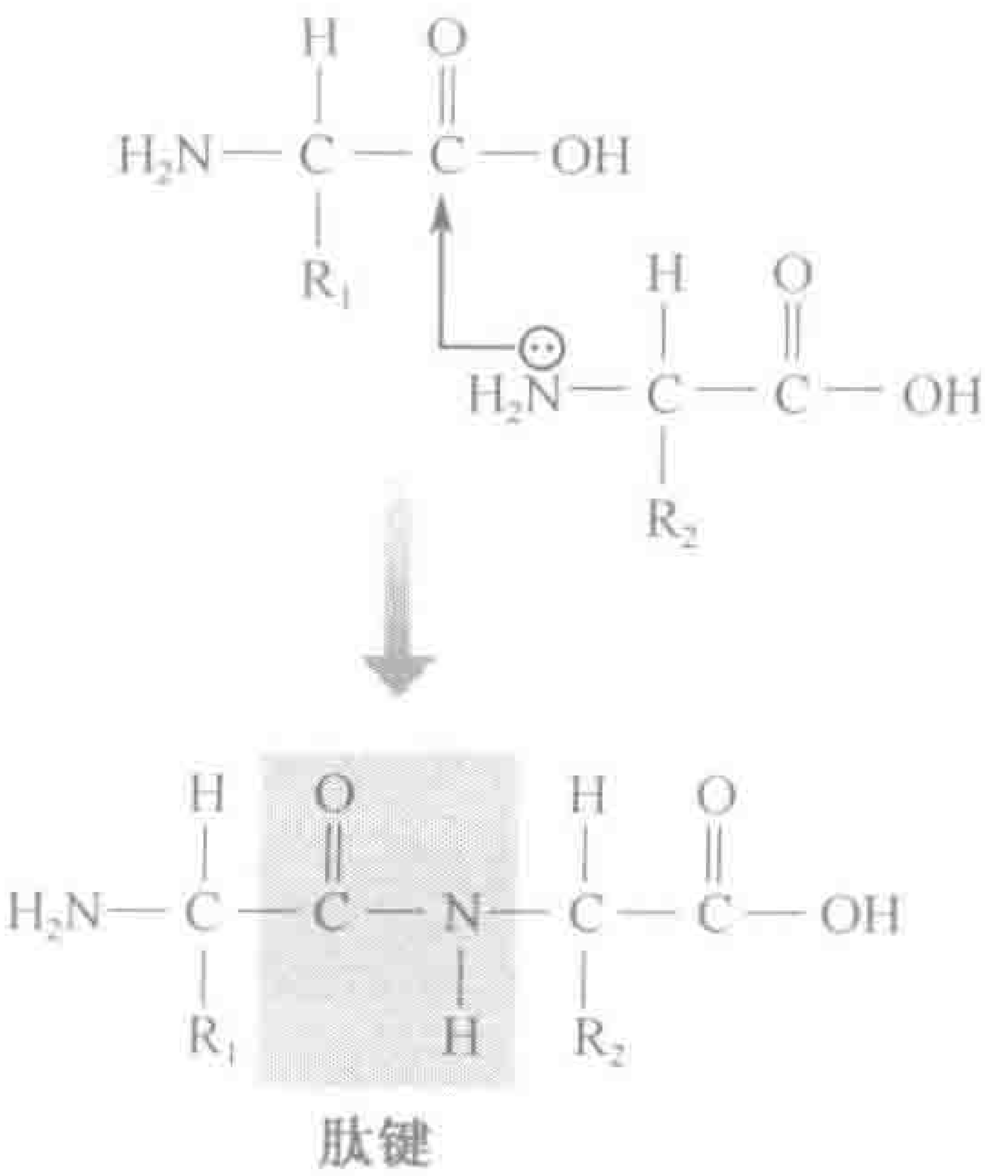


图 1.21　多肽通过连续的氨基酸之间的肽键形成而合成

翻译的一种模式设想为最初 40S 核糖体亚基通过特异性与帽结合的蛋白质的参与来识别 5′帽。然后它沿着 mRNA 移动直到遇到起始密码子，起始密码子几乎总是 AUG，特定的蛋氨酸（少数情况下已知使用 ACG、CUG 或 GUG 来代替）。通常遇到的第一个 AUG 为起始密码子，但不总是这样。然而 AUG 只有嵌入一适宜的**起始密码子识别序列**（initiation codon recognition sequence）时才能够作为一个起始密码子被有效识别，最理想的起始密码子识别序列为 GCCPuCCAUGG。此序列中最重要的决定因素是 AUG 密码子之后的 G，以及在它之前三个核苷酸的嘌呤（Pu），Pu 最好是 A（Kozak，1996）。随后，通

过缩合反应，逐个的氨基酸被整合到不断增长的多肽链中：为了整合，进入的氨基酸的氨基与最后氨基酸的羧基反应，在连续的残基之间形成一个肽键（图 1.21）。这是由肽键转移酶催化的，该酶存在于大核糖体亚基的 RNA 组分中。

1.5.2　遗传密码是简并的，并非完全通用的密码

遗传密码是一个由三个字母组成的密码。在一个密码子三个碱基位置的每一处都有四种可能的碱基用于选择。所以，有 $(4)^3=64$ 种可能的密码子，但只有 20 种主要类型氨基酸。因此，遗传密码是一个**简并密码**（degenerate code）：每个氨基酸平均大约由三个不同的密码子所表示，尽管有些氨基酸，如亮氨酸和丝氨酸由六个密码子表示而其他氨基酸由更少的密码子表示（图 1.22）。

密码子	氨基酸	密码子	氨基酸	密码子	氨基酸	密码子	氨基酸
AAA	Lys	CAA	Gln	GAA	Glu	UAA	终止密码子
AAG		CAG		GAG		UAG	
AAC	Asn	CAC	His	GAC	Asp	UAC	Tyr
AAU		CAU		GAU		UAU	
ACA	Thr	CCA	Pro	GCA	Ala	UCA	Ser
ACG		CCG		GCG		UCG	
ACC		CCC		GCC		UCC	
ACU		CCU		GCU		UCU	
AGA	Arg(N) / 终止密码子（线粒体）	CGA	Arg	GGA	Gly	UGA	终止密码子（细胞核） / **Trp**（线粒体）
AGG		CGG		GGG		UGG	Trp
AGC	Ser	CGC		GGC		UGC	Cys
AGU		CGU		GGU		UGU	
AUA	Ile(N) / **Met**（线粒体）	CUA	Leu	GUA	Val	UUA	Leu
AUG	Met	CUG		GUG		UUG	
AUC	Ile	CUC		GUC		UUC	Phe
AUU		CUU		GUU		UUU	

图 1.22　核遗传密码与线粒体遗传密码相似但不相同

四个密码子（以阴影显示）在哺乳细胞的细胞核与线粒体中解释不同，在线粒体中的解释以黑体表示。因此，线粒体密码有四个而不是三个终止密码子（UAA、UAG、AGA、AGG），两个而不是一个色氨酸密码子（UGA、UGG），四个而不是六个精氨酸密码子（CGA、CGC、CGG、CGU），两个而不是一个蛋氨酸密码子（AUA、AUG）以及两个而不是三个异亮氨酸密码子（AUC、AUU）。**注**：(1) 遗传密码的简并性大多数通常涉及密码子的第三个碱基。有时任何一个碱基可被替代（GGN＝甘氨酸，CCN＝脯氨酸，N 为任何一个碱基）。在其他情况，任何一个嘌呤（Pu）或任何一个嘧啶（Py）可被替代（AAPu＝赖氨酸，AAPy＝天冬酰胺等）。(2) 一些 mRNA 中信号能够造成终止密码子的可选择性解释，例如 UGA 可特定第 21 种氨基酸，硒代半胱氨酸，而 UAG 可特定谷氨酰胺或第 22 种氨基酸，吡咯赖氨酸（Atkins and Gesteland，2002）。

虽然有 64 种密码子，但是具有不同反密码子的 tRNA 分子的相应数目比较少：仅有 30 多种胞质 tRNA 和 22 种线粒体 tRNA。在胞质和线粒体核糖体上，所有 64 种密码子的解译都是有可能的，因为当正常的碱基配对法则遇到密码子-反密码子识别时变得宽松了。**“摆动假说”**（wobble hypothesis）表明，在一个密码子的前两个碱基位置上，密码子与反密码子的配对遵循正常的 A-U 和 G-C 法则，但在第三个位置上发生了特殊的“摆动”，也可使用 G-U 碱基配对（表 1.6）。

表 1.6 密码子-反密码子配对在密码子的第三个碱基位置允许宽松的碱基配对（摆动）

tRNA 反密码子 5′端碱基	mRNA 密码子 3′端识别的碱基
A	只有 U
C	只有 G
G（或 I）*	C 或 U
U	A 或 G

* I＝次黄嘌呤核苷，腺嘌呤的一种翻译后修饰形式。更多的关于次黄嘌呤核苷和胞质 tRNA 密码子-反密码子识别的细节见框 9.4。

通常将遗传密码描述成一个**通用密码**（universal code），意味着在所有生命形式中均使用相同的密码。这由于（下列原因）造成的密码子选择性解译而并非绝对真实：

▶ **细胞器遗传密码的进化趋异**。线粒体和叶绿体也有有限的蛋白质合成能力。在进化过程中，这些细胞器采用了与核基因稍有不同的遗传密码。因此，例如核编码的 mRNA 的翻译持续进行，直到遇到一个**终止密码子**（termination codon），它是三种可能的密码子（UAA、UAG、UGA）之一，而在哺乳动物线粒体中有四种可能的密码子——UAA、UAG、AGA 和 AGG（图 1.22）。

▶ **密码子的邻近依赖性重新定义**。在一些 mRNA 中，包括几种类型核编码的 mRNA，信号造成一些密码子的重新定义。例如，在许多细胞中（包括人类细胞），一些核编码的 mRNA 可选择性地将 UGA 解译为第 21 种氨基酸，**硒代胱氨酸**（selenocysteine），而 UAG 可被选择性解译为谷氨酰胺（Atkins and Gesteland，2002）。

因此，初级翻译产物的骨架在一个末端将是具有一个游离氨基的蛋氨酸（**N 端**，N-terminal end），而在另一个末端将是具有一个游离羧基的氨基酸（**C 端**，C-terminal end）。注意，尽管密码子是在一特定的**翻译读框**（translational reading frame）内进行翻译，但在真核细胞中偶尔可见重叠基因，它们使用了不同的翻译可读框（例子见图 9.3）。

控制翻译的主要步骤是核糖体结合。为了翻译，除了 5′帽之外，5′UTR（通常<200bp）和 3′UTR 均在 mRNA 募集中发挥关键作用。已经描述了参与此过程的几个顺式作用元件的特性，此外还鉴定了与这些元件结合的反式作用因子。为增强翻译，5′和 3′UTR 序列有可能相互作用。3′UTR 在翻译调控中具有关键作用，而且控制翻译、mRNA 稳定性及定位的信号均发现于此区域内（Wickens *et al.*，1997；也见节 10.2.6）。

1.5.3 翻译后修饰包括某些氨基酸的化学修饰和多肽切割

初级翻译产物通常经历各种修饰反应，包括在翻译水平和翻译后水平添加的化学基团，这些基团共价地附着于多肽链。这可以涉及单个氨基酸侧链的简单化学修饰（羟化、磷酸化等）或者添加不同类型碳水化合物或脂类基团（表 1.7）。

表 1.7 多肽修饰的主要类型

修饰的类型（添加的基团）	靶氨基酸	注释
磷酸化（PO_4^-）	酪氨酸，丝氨酸，苏氨酸	由特定激酶完成，可由磷酸酶使其逆转
甲基化（CH_3）	赖氨酸	由甲基化酶完成，去甲基化酶使其还原
羟化（OH）	脯氨酸，赖氨酸，天冬氨酸	羟脯氨酸和羟赖氨酸特别常见于胶原中
乙酰化（CH_3CO）	赖氨酸	由乙酰化酶完成，去乙酰化酶使其还原
羧化（COOH）	谷氨酰胺	由γ羧化酶完成
乙酰化（CH_3CO）	赖氨酸	由乙酰化酶完成，去乙酰化酶使其还原
N-糖基化（复杂碳水化合物）	天冬酰胺，通常位于天冬酰胺-X-丝氨酸/苏氨酸序列中	最初发生于内质网；X为脯氨酸之外的任何氨基酸
O-糖基化（复杂碳水化合物）	丝氨酸，苏氨酸，羟赖氨酸	发生于高尔基体，不如N-糖基化常见
GPI（糖脂）	C端的天冬氨酸	用于将蛋白质锚定在质膜的外层
豆蔻酰化/十四烷酰化（C_{14}脂肪酰基团）	N端的甘氨酸（见正文）	用作膜锚定物
棕榈酰化/十六烷酰化（C_{16}脂肪酰基团）	形成S-棕榈酰基连接的半胱氨酸	用作膜锚定物
法尼基化（C_{15}异戊二烯基团）	C端的半胱氨酸（见正文）	用作膜锚定物
牻牛儿基丙酮化（C_{20}异戊二烯基团）	C端的半胱氨酸（见正文）	用作膜锚定物

通过添加碳水化合物基团进行蛋白质修饰

糖蛋白（glycoprotein）含有共价附加于某些氨基酸侧链的寡糖。细胞液中几乎没有蛋白质是糖基化的，即已经附加了碳水化合物，而那些携带一单个糖残基——N乙酰葡萄糖胺的蛋白质则共价连接一个丝氨酸或苏氨酸残基。相反，那些从细胞分泌或者输出至溶酶体、高尔基体或细胞膜的蛋白质是糖基化的。糖蛋白的寡糖成分被大量地预先形成并整个地添加至多肽。两种主要类型的糖基化得到公认：

- **N糖基化**（N-glycosylation）：大多数情况下，一个常见的寡糖序列初始转运到在内质网内的一个天冬酰胺残基的侧链 NH_2 基团上（ER；表1.7）。随后（在高尔基体内发生）。残基修整以及用不同单糖置换。
- **O糖基化**（O-glycosylation）见表1.7。

蛋白多糖（proteoglycan）是附加了葡萄糖胺聚糖的蛋白质，葡萄糖胺聚糖通常含有葡萄糖胺或半乳糖胺的二糖重复单位。最具特性的蛋白多糖是细胞外基质的组成成分。

通过添加脂类基团进行蛋白质修饰

一些蛋白质，特别是膜蛋白，是通过添加脂肪酰基或异戊二烯基进行修饰的，脂肪酰基或异戊二烯基一般用作膜锚定物。脂肪酰基的例子包括**豆蔻酰基**（myristoyl group），它是一个 C_{14} 脂质，添加至位于最远 N 端的甘氨酸残基上，使得修饰的蛋白质能够与膜受体或膜的脂质双层相互作用。另一个用作膜锚定物的脂肪酰基是 C_{16} **棕榈酰基**（palmitoyl group），它被添加至半胱氨酸残基的 S 原子上。

异戊二烯基（prenyl group）典型地添加至接近 C 端的半胱氨酸残基上，包括**法尼基**（farnesyl group）（C_{15}）和**牻牛儿基丙酮基**（geranylgeranyl group）（C_{20}）。许多参与信号转导和蛋白质靶向的蛋白质在它们的 C 端含有一个法尼基单位或一个牻儿基丙酮单位。

蛋白质与细胞膜外层的锚定使用了一个不同的机制：**糖化磷脂酰肌醇**（glycosylphosphatidyl inositol，GPI）基团的附着。这个复杂的糖脂基团含有一个用作膜锚定物的脂肪酰基，可连续地连接到一个磷酸甘油单位和一个寡糖单位，并最后通过一个磷酸乙醇胺单位连接至蛋白质的 C 端。除了 GPI 锚定物之外，整个蛋白质均位于细胞外的空间。

翻译后切割

初级翻译产物也可以经过内部切割产生一个较小的成熟产物。有时候，开始的蛋氨酸从初级翻译产物上切割，如在 β 珠蛋白的合成过程中（图 1.19）。对于许多蛋白质（包括血浆蛋白、多肽激素、神经肽、生长因子等）的成熟来说，可观察到更多的多肽切割。就像在下一节描述的那样，切割信号序列常用于标签蛋白质，这些蛋白质派定被送至某个特定的位置。另外，在一些情况下，由于对一个大的前体多肽的蛋白水解作用的结果，所以一单个 mRNA 分子可以特化不只一种功能的多肽链（图 1.23）。

1.5.4　蛋白质分泌和细胞内输出是由特异的定位信号和化学修饰所控制

在线粒体核糖体合成的蛋白质需要在线粒体内发挥作用。然而，在胞质核糖体合成的大批蛋白质具有不同功能，它们需要从合成它们的细胞中分泌（就像激素和其他细胞间信号分子）或者输出至特定的细胞内结构，例如细胞核（组蛋白、DNA 和 RNA 聚合酶、转录因子、RNA 加工蛋白等），线粒体（线粒体核糖体蛋白、一些呼吸链组分等），过氧化物酶体等。为了实现这一目的，一个特异的**定位信号**（localization signal）必须嵌入至多肽的结构中，以便它能够被输送到正确的位置。定位信号通常表现为一个短的多肽序列形式。但通常并非总是，此序列构成了一个所谓的**信号序列**（signal sequence）（或**前导序列**，leader sequence），一旦完成分类过程，该序列就被一专门的**信号肽酶**（signal peptidase）从蛋白质中除掉。

用于输出至内质网和细胞外空间的信号

对于分泌性蛋白质来说，信号肽由 N 端最初 20 个左右的氨基酸组成，并且总是包括大量的疏水氨基酸（表 1.8）。信号序列由一个**信号识别颗粒**（signal recognition par-

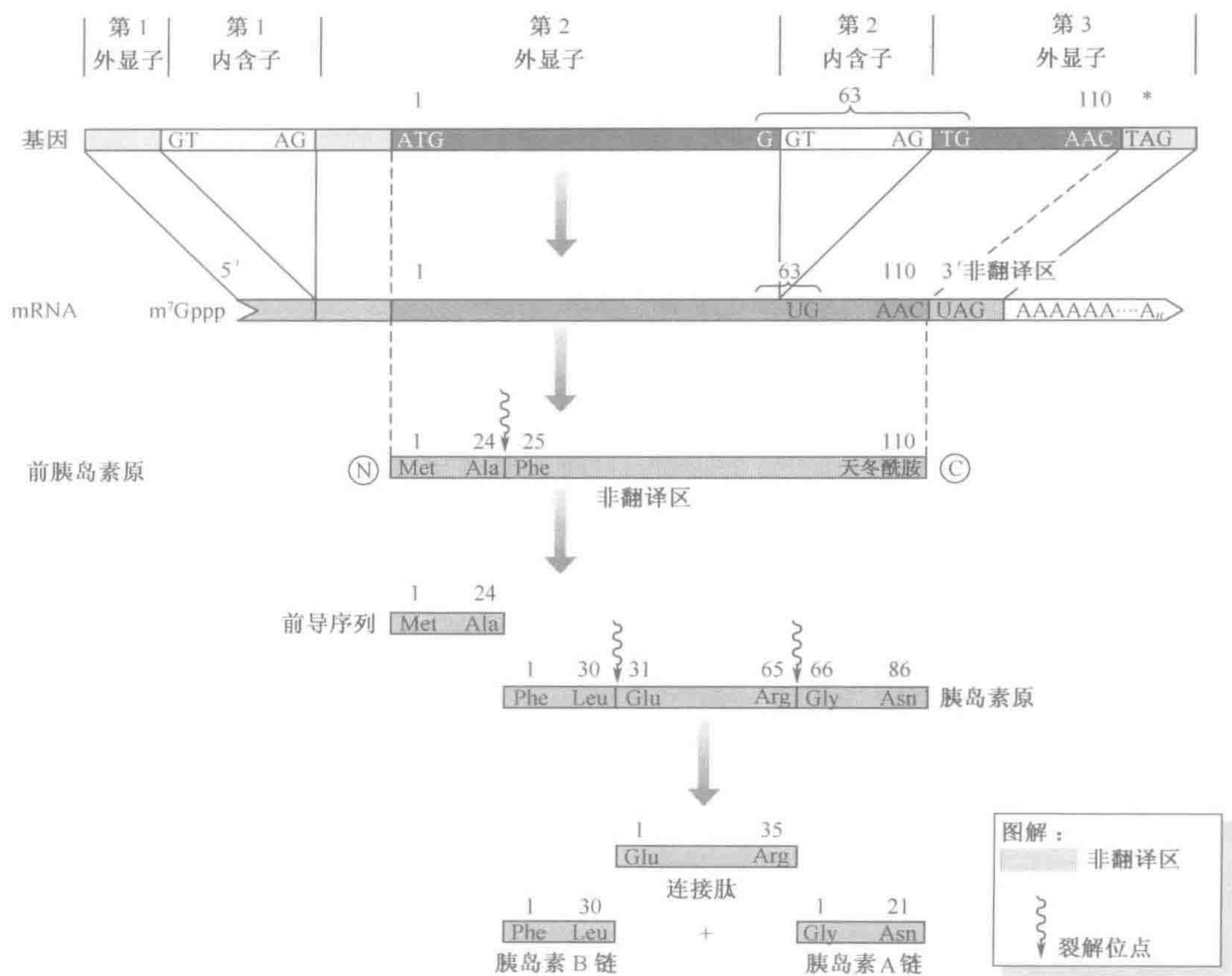


图 1.23　胰岛素合成涉及多肽前体的多个翻译后切割

第一个内含子中断 5′非翻译区；第二个内含子中断密码子 63 的第一位与第二位碱基，并被归类为一个 1 位相内含子（其他内含子位相见框 12.2 及图 1.19）。初级翻译产物，前胰岛素原具有一个 24 个氨基酸的**前导序列**（lead sequence），此序列是蛋白质通过细胞膜所必需的，并在此后被除去。胰岛素原前体含有一个中心节段——**连接肽**（connecting peptide），它对于维持 A 链和 B 链节段的构象以便它们能够形成二硫桥是很重要的（图 1.25）。

ticle，SRP）引导至内质网。信号识别颗粒是一种小的胞质 RNA（7SL RNA）和六个特定蛋白质组成的 RNA-蛋白质复合物。SRP 复合物可与不断增长的多肽链以及核糖体结合并指导他们到达位于粗面内质网的细胞溶质侧表面的 SRP 受体蛋白。此后，多肽能够进入内质网腔，被派定从细胞中输出。除非存在额外的、终止转运过程的疏水性片段，如**跨膜蛋白**（transmembrane protein）。

其他信号

与内质网信号一样，一个 N 端信号序列是跨越线粒体膜所必需的，随后该序列被切割。一般地，除了许多疏水氨基酸之外，线粒体信号肽具有几个带正电荷的氨基酸，通常间隔大约四个氨基酸。这个结构将形成一个兼性 α 螺旋，此螺旋结构在一面为带电氨基酸而另一面为疏水氨基酸（见下文及图 1.24A）。

核定位信号（nuclear localization signal）可位于多肽序列内的任何位置，一般由一串4～8个带正电的氨基酸以及相邻的脯氨酸残基组成。然而，这个信号通常是两部分，带正电的氨基酸可见于两个由2～4个残基组成的区块，其间隔约10个氨基酸（block）（表1.8）。值得注意的是，有些核蛋白本身缺乏任何核定位序列，但是借助于其他具有适当信号的核蛋白可运送至细胞核内。

表1.8 蛋白质定位序列的例子

蛋白质的目的地	位置和信号种类	例子
内质网以及从细胞分泌	20个左右氨基酸的N端肽；非常疏水的	人类胰岛素——24个氨基酸，高度疏水的信号肽：N-蛋-丙-亮-色-蛋-精-亮-亮-脯-亮-亮-丙-亮-亮-丙-亮-色-甘-脯-天冬-脯-丙-丙-丙
线粒体	N端肽；带正电的残基位于一面而疏水残基位于另一面的α螺旋	人类线粒体乙醛脱氢酶N端17个氨基酸：N-蛋-亮-**精**-丙-丙-丙-**精**-苯丙-甘-脯-**精**-亮-甘-**精**-**精**-亮-亮
细胞核	氨基酸内部序列；通常为一串碱性氨基酸和脯氨酸；可以是有两个部分的	SV40 T抗原——连续的：脯-脯-**赖**-**赖**-**赖**-精-**赖**-缬
溶酶体	添加甘露糖6-磷酸残基	p53——分为两部分的：**赖**-**精**-丙-亮-脯-天冬酰胺-天冬酰胺-苏-丝-丝-丝-脯-谷氨酰胺-脯-**赖**-**赖**-**赖**

溶酶体蛋白（lysosomal protein）通过添加甘露糖-6-磷酸残基靶向到达溶酶体，该残基添加至高尔基体的**顺式区室**（*cis*-compartment）并可被高尔基体**反式区室**（*trans*-compartment）的受体蛋白识别。

1.5.5 蛋白质结构是高度变化的，不容易从氨基酸序列预测

蛋白质是由一条或多条多肽组成，每条多肽都可经历翻译后修饰。它们可与特定的**辅助因子**（co-factor）（例如二价阳离子如Ca^{2+}、Fe^{2+}、Cu^{2+}、Zn^{2+}，或者功能性酶活性所必需的小分子如NAD^{+}）或者**配体**（ligand）（一个蛋白质特异性结合的任何分子）相互作用，每一个可对蛋白质的构象产生强有力的影响。蛋白质至少已分为四种不同级别的结构组织（表1.9）。

表1.9 蛋白质结构的级别

级别	定　义	注　释
一级	一条多肽中氨基酸的线性序列	长度变化巨大，从一个小的肽分子到几千个氨基酸长
二级	一个多肽骨架遵循的途径	可以局部变化，例如像α螺旋或β折叠
三级	一条多肽的全部三维结构	变化巨大，例如球状的，杆状的，管，线圈，片状等
四级	一多聚蛋白质的全部结构	通常由二硫桥以及与配体结合等稳定

在一单个多肽链内，不同残基之间有大量的氢键形成机会；不考虑侧链，一个肽键

羰基（CO）基团的氧原子能够与另一个肽键氨基（NH）基团的氢原子形成氢键。由一单个多肽紧密相邻的氨基酸残基之间氢键形成所确定的基本结构单位包括：

- **α 螺旋**（α-helix）。这涉及一个刚性的圆柱体的形成。结构由一个肽键的羰基氧与相距四个氨基酸的肽键的氨基氢原子之间的氢键形成所控制（图 1.24）。注意，转录因子的 DNA 结合结构域通常为 α 螺旋的（节 10.2.4）。**兼性 α 螺旋**（amphipathic α-helix）在一侧表面具有带电残基而另一面为疏水残基（图 1.24）。具有一个非极性侧链重复排列的相同的 α 螺旋能够彼此缠绕，形成一个特殊的稳定结构，称为**卷曲螺旋**（coiled coil）。长的杆状卷曲螺旋常见于许多纤维蛋白，如皮肤、头发、指甲的 α-角质纤维或者血凝块的纤维蛋白原。
- **β 折叠**（β-pleated sheet）。它是以同一多肽链平行部分（实际上常为反向平行）的反向肽键之间氢键形成为特征（图 1.24）。β 折叠形成大多数但不是全部的球状蛋白质的核心。
- **β 转角**（β-turn）。一条多肽第 n 个氨基酸残基的肽键羰基（CO）与第 $n+3$ 个氨基酸残基的肽键 NH 基团之间的氢键形成的结果是一个发夹状转角。通过突然地使多肽转向相反方向，可获得紧凑的球状外形。β 转角如此命名是因为它们通常也连接 β 折叠的反向平行链（图 1.24B）。

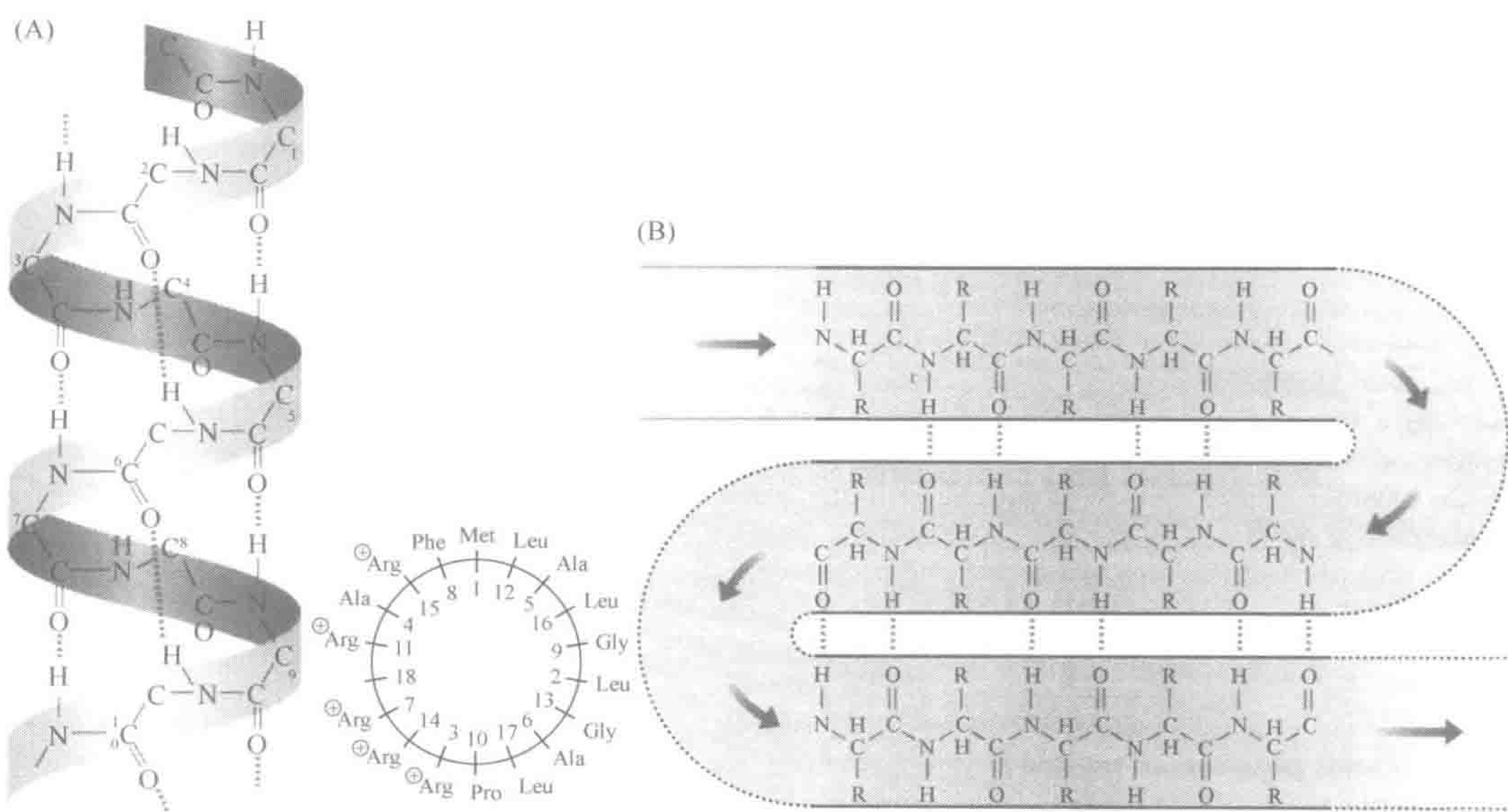


图 1.24　多肽二级结构区域通常受链内氢键形成控制

（A）**一个 α 螺旋的结构**。**左**：为了清晰，仅显示多肽的骨架。每个肽键的羰基（CO）氧原子与远离的第四个氨基酸肽键氨基（NH）的氢原子形成氢键，因此螺旋每圈含有 3.6 个氨基酸。**注**：出于清晰的目的，省略了一些键。**右**：在一兼性 α 螺旋中，带电的氨基酸与疏水性氨基酸位于不同的表面。显示的序列是线粒体乙醛脱氢酶 17 个氨基酸长的信号肽序列（表 1.8）。（B）**一个 β 折叠的结构**。**注**：此氢键形成发生于多肽骨架邻近平行部分的肽键的羰基氧原子和氨基氢原子之间。例子显示了多肽骨架反向平行部分之间键形成的一种情况（反向平行 β 折叠），而反向平行部分之间强迫性方向的突然改变通常利用 β 转角完成（见正文）。箭头标示了从 N 端到 C 端的方向。**注**：也常见同一方向走行的邻近部分形成的 β 折叠。

由上述结构单元联合组成的更复杂的结构基序构成了**蛋白质结构域**（protein domain），即由一个蛋白质的一级结构向后折叠以便二级结构元件能够彼此相邻堆积而形成的紧凑区域。这些结构域通常是参与和其他分子结合的功能单位。另外，同一多肽链内或不同的多肽链上存在的成对的半胱氨酸残基的巯基（—SH）基团之间常形成共价**二硫桥**（disulfide bridge）（图 1.25）。

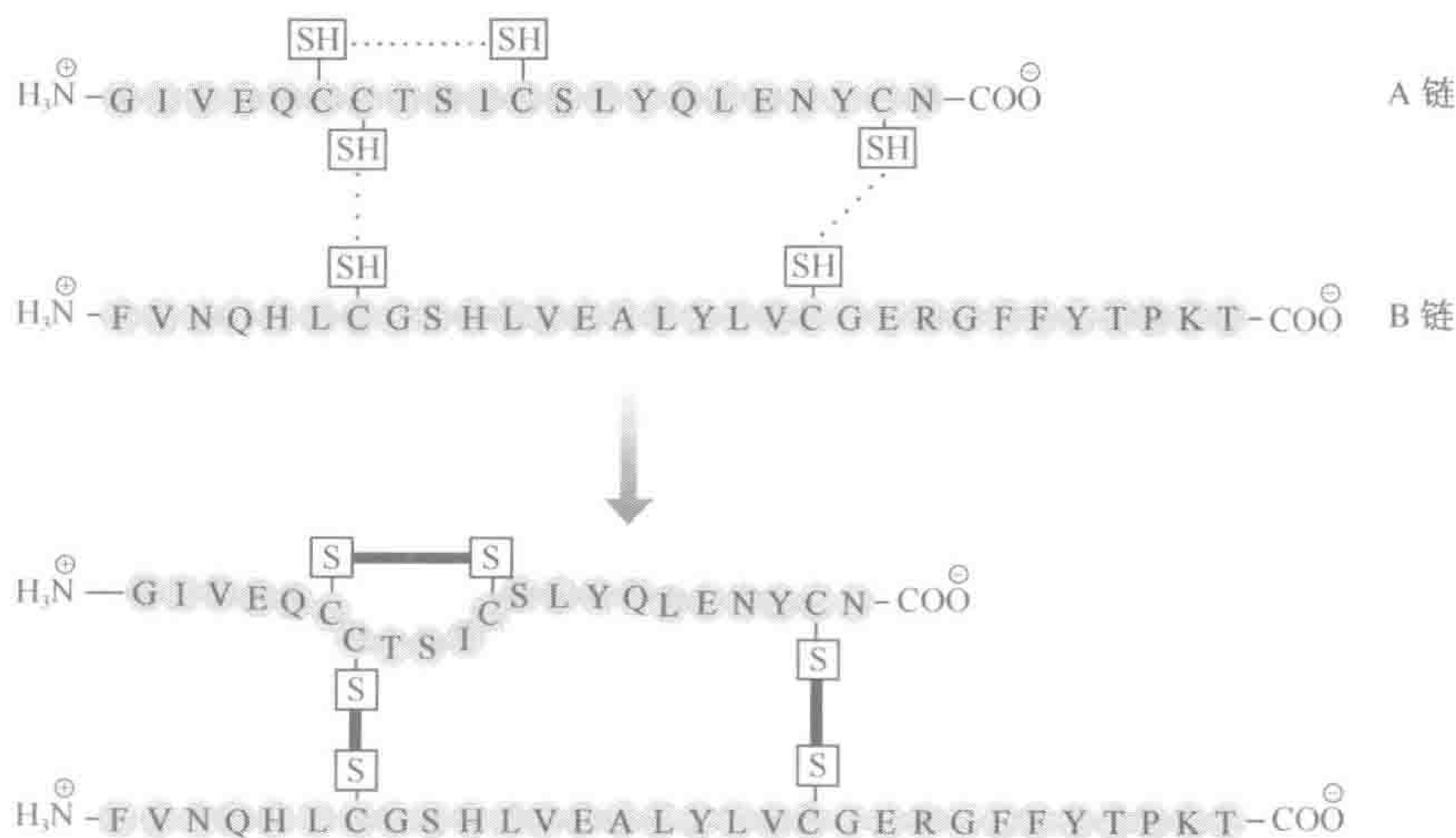


图 1.25　人类胰岛素中链内与链间二硫桥

二硫桥（-S-S-）是通过 A 链第六个和第十一个半胱氨酸残基相对的巯基（—SH）基团之间，或者不同链标明的残基之间的一个缩合反应形成的。

很明显，蛋白质的三级或四级结构由一级氨基酸序列所决定。然而，尽管二级结构基序如 α 螺旋、β 折叠和 β 转角能够通过分析初级序列而预测，但是目前仍无法准确地预测总的三维结构。除了单一多肽的结构复杂性之外，许多蛋白质组织为多个多肽亚单位的复杂聚集物。

（李晓明　译）

进一步阅读

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2001) *Molecular Biology of the Cell*, 4th Edn. Garland Publishing, New York.

Bell SP, Dutta A (2002) DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71**, 307–331.

Big Picture Book of Viruses at: http://www.virology.net/Big_Virology/BVHomePage.html

Braden C, Tooze J (1999) *Introduction to Protein Structure*, 2nd Edn. Garland Science, New York.

Brow DA (2002) Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.* **36**, 333–360.

Carradine CR, Drew HR (1997) *Understanding DNA. The molecule and how it works.* Academic Press, London.

Lodish H, Baltimore D, Berk A, Zipursky L, Matsudaira P, Darnell J (1995) *Molecular Cell Biology*, 3rd Edn. Scientific American Books, New York.

参考文献

Atkins JF, Gesteland R (2002) The 22nd amino acid. *Science* **296**, 1409–1410.

Berget SM (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414.

Friedberg EC, Feaver WJ, Gerlach VL (2000) The many faces of DNA polymerases: strategies for mutagenesis and for mutational avoidance. *Proc. Natl Acad. Sci. USA* **97**, 5681–5883.

Friedberg EC, Wagner R, Radman M (2002) Specialised DNA polymerases, cellular survival, and the genesis of mutations. *Science* **296**, 1627–1630.

Hastings ML, Krainer AR (2001) Pre-mRNA splicing in the new millennium. *Curr. Opin. Genet. Dev.* **13**, 302–309.

Kozak M (1996) Interpreting cDNA sequences: some insights from studies on translation. *Mamm. Genome* **7**, 563–574.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1113.

Lim LP, Burge CP (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA* **98**, 11193–11198.

Staley JP, Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs and things. *Cell* **92**, 315–326.

Tarn WY, Steitz JA (1997) Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge. *Trends Biochem. Sci.* **22**, 132–137.

Wickens M, Anderson P, Jackson RJ (1997) Life and death in the cytoplasm: messages from the 3′ end. *Curr. Biol.* **7**, 220–232.

第2章　染色体结构和功能

本章内容

DNA在一定的环境中起作用。人类DNA构成**染色体**（chromosome），并在细胞中执行功能，而这些细胞是来自其他分裂的细胞。当然，细胞分裂的过程是**细胞周期**（cell cycle）的一小部分，在此过程中染色体及其组成的DNA分子需要形成和其自身完全一致的拷贝，然后分离到子细胞中去。这一过程需要非常精细的协调以保证子细胞接受到正确的染色体组。细胞分裂的普遍形式——**有丝分裂**（mitosis）——子细胞与亲代细胞具有相同的染色体数目和类型。此外，细胞分裂的一种特殊形式——**减数分裂**（meiosis）——发生在睾丸和卵巢的特定细胞，并分别产生精子和卵细胞。

2.1　倍性和细胞周期

在任何有核细胞中染色体的数目不同，**染色体组**（chromosome set）和与之相关的DNA含量分别用n和C来标明。就人类而言，$n=23$，$C\approx3.5$pg（3.5×10^{-12}g）。细胞可能因其所含有染色体组（**倍性**，ploidy）的不同而拷贝数不同。精子和卵细胞携带单一染色体组而称为**单倍体**（haploid，n条染色体，DNA含量是C）。然而，大多数人类细胞携带染色体组的两个拷贝，是**二倍体**（diploid，$2n$条染色体，DNA含量$=2C$）。几乎所有的哺乳动物都是二倍体（红兔鼠是一个罕见的例外——见Gallardo *et al.*，1999），但在其他有机体中有许多正常的单倍体、**四倍体**（tetraploid，$4n$）或**多倍体**（polyploid，$>4n$）物种的例子。**三倍体**（triploidy，$3n$）因其在减数分裂时出现的问题而较少见。

人体的二倍体细胞最终全部都来自单一的二倍体细胞，即**合子**（zygote），通过有

丝分裂重复循环而来的。每一个分裂循环可总结为一个**细胞周期**（cell cycle）（图 2.1），它包括一个短的细胞分裂阶段 **M 期**（M phase）（图 2.7）和一长间隙的**间期**（interphase）。间期可以分为 **S 期**（**S phase**）（DNA 合成期），**G_1 期**（G_1 phase）（M 期与 S 期之间的间隙）和 **G_2 期**（G_2 phase）（S 期与 M 期之间的间隙）。从有丝分裂的后期一直到在 S 期 DNA 复制，一个二倍体细胞中的一条染色体含有一单个的 DNA 双螺旋而且 DNA 总量是 $2C$。G_1 期是细胞的正常状态而且是非分裂细胞的长期、终末的状态。只要细胞进行有丝分裂它们即进入 S 期；非分裂细胞仍停留在一个修饰的 G_1 期阶段，有时称为 **G_0 期**（G_0 phase）。细胞周期图给我们这样一个印象，即所发生的全部有意义的行为都在 S 期和 M 期中，但这是一个错觉。细胞在其生命的大部分时间都处于 G_0 期或 G_1 期，而基因组所做的大部分工作正是在这期间进行的。

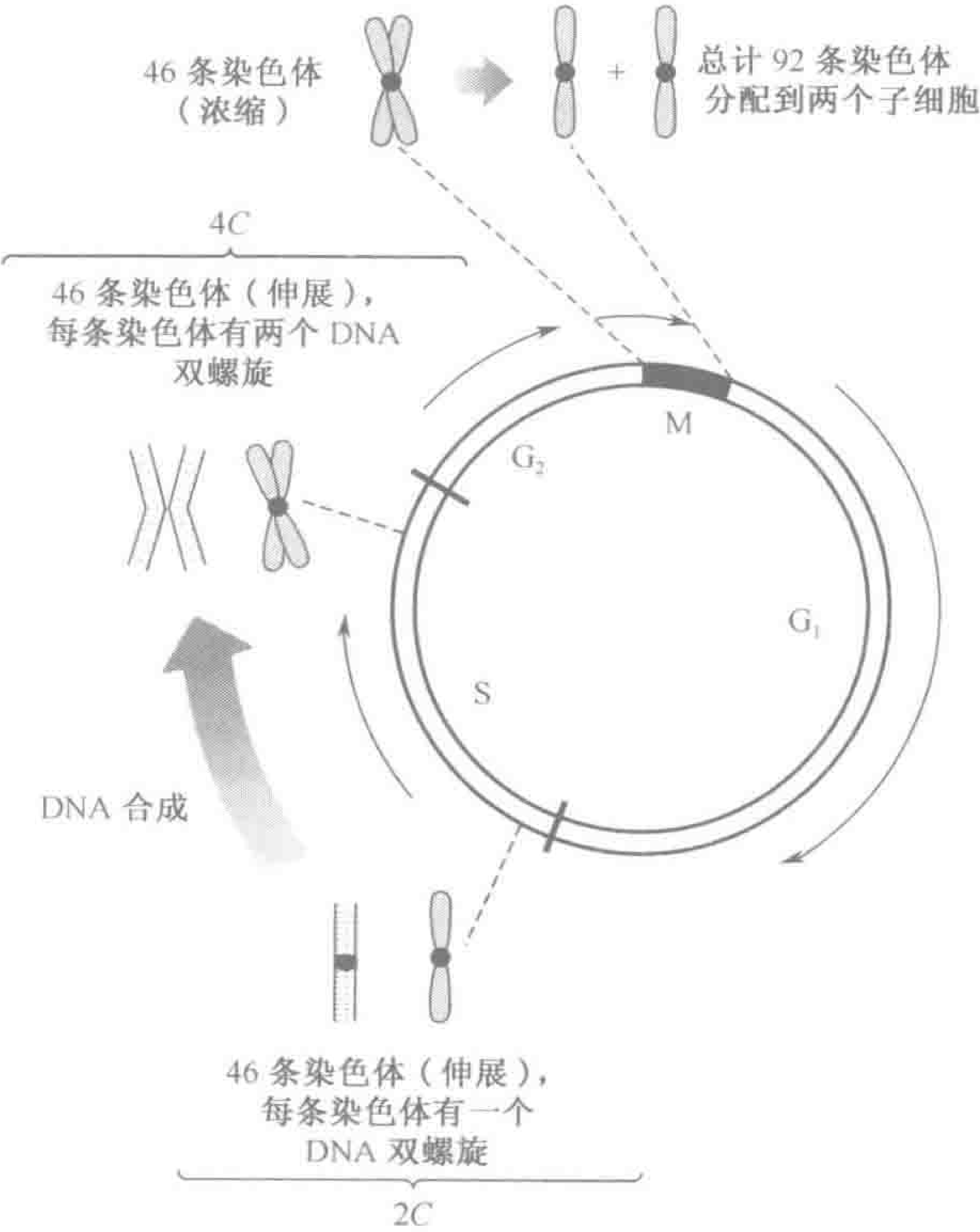


图 2.1 在细胞周期中人类染色体的 DNA 含量

间期包括 G_1 + S+ G_2，从有丝分裂的后期起一直到 S 期 DNA 复制前染色体含有一个 DNA 双螺旋。从 S 期 DNA 复制完成开始一直到有丝分裂中期结束，每条染色体含有两个姐妹染色单体，每条染色单体含有一个 DNA 双链，形成每条染色体的两个双螺旋。一个二倍体细胞的 DNA 含量在 S 期之前是 $2C$（单倍体细胞 DNA 含量的二倍），而在 S 期与有丝分裂之间是 $4C$。

二倍体细胞的一亚群形成**生殖系**（germ line）（图 2.9），在卵巢和睾丸中产生特殊的二倍体细胞，这些二倍体细胞可以通过减数分裂产生单倍体**配子**（gamete，精子和卵）。在人类（$n=23$）每个配子中含 22 条**常染色体**（autosome chromosome，非性染色体）加 1 条性染色体。在卵细胞中的性染色体总是 X；在精子中，它可是 X 或 Y。在

受精之后，合子是二倍体（$2n$），其染色体组成是 46，XX 或 46，XY（图 2.2）。除了生殖系，身体的其他细胞称为**体细胞**（somatic cell）。人类体细胞通常是二倍体，但其中一些细胞没有细胞核和任何染色体，它们被称为**缺倍体**（nulliploid），而其他的则有一些是自然的多倍体，是由于 DNA 复制了多次而无细胞分裂的结果（节 3.1.4）。

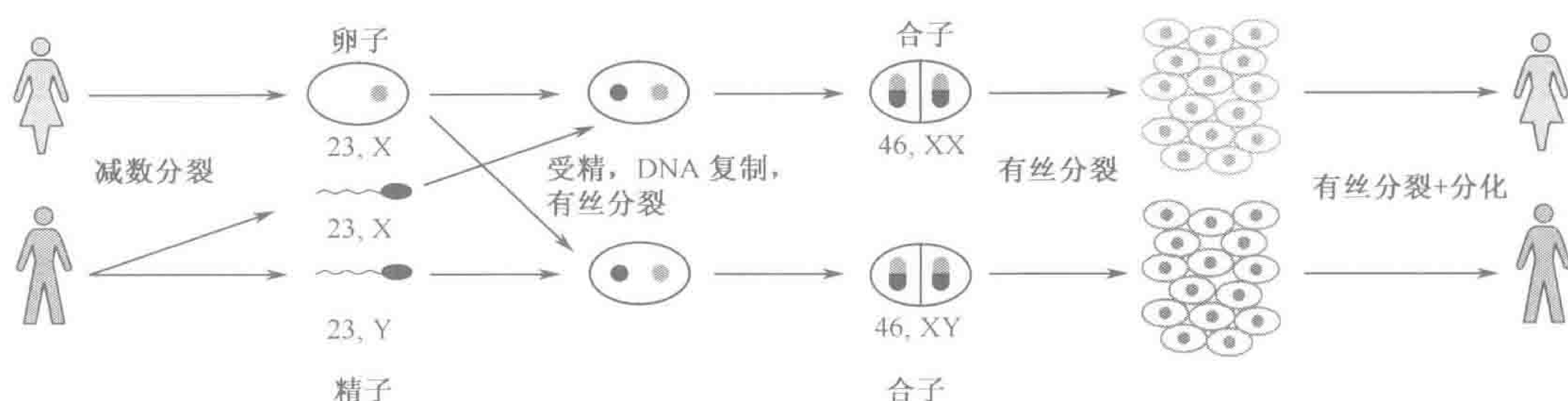


图 2.2　从染色体角度看人类生命

由二倍体前体细胞通过减数分裂产生单倍体精子和卵细胞（图 2.9）。在受精卵中精子和卵细胞染色体的起初形式分别是雄原核和雌原核，它们在第一次有丝分裂时才融合在一起。

2.2　染色体的结构和功能

在显微镜下所见到的和教科书中所描述的染色体很容易使人产生误解，因为它们所描述的是在细胞周期中所发生一个短暂的、不寻常的状态，是在细胞准备进行细胞分裂（中期）时的 M 期的一部分。**中期染色体**（metaphase chromosome）和**前中期染色体**（prometaphase chromosome）处于高度凝缩状态具有一个不寻常的两个染色单体结构，因为在此阶段为细胞分裂作准备的 DNA 已经复制。中期染色体通过紧密地包装成简洁的束状结构，它足够大至可用光学显微镜观察，但是它们此时并不表达，因为极端紧密的包装确保基因的关闭。

细胞分裂的过程，以其自身正确和错误地包装或基因组分割会导致重要的医学后果而令人惊叹（节 2.5）。然而记住这一点是重要的，那就是细胞周期的很大部分染色体具有十分不同的外形。在全长的间期阶段，染色体更加伸展，所以比在图 2.14 所见的中期染色体更为扩散。重要的是**间期染色体**（interphase chromosome）的组成只有单条染色单体和一个 DNA 双螺旋，而这伸展的结构允许基因表达。

作为功能性细胞器，真核生物的染色体似乎仅需要三类 DNA 序列元件：着丝粒、端粒和复制起始点。这一简单要求已经被酵母人工染色体的成功构建所证实：外源大片段的 DNA 序列连接到特化为一个功能的着丝粒、两个端粒和一个复制起始点的短序列时，就会作为自主性的染色体（图 5.17）。最近**哺乳动物人工染色体**（mammalian artificial chromosome）以相似的原理而被构建（Huxley，1997；Schindelhauer，1999）。

2.2.1　DNA 组装成染色体需要多层次 DNA 的折叠

在细胞中每一条染色体的结构都是高度有序的（Manuelidis，1990）。即使在间期

细胞核的 2nm DNA 双螺旋构成至少经受两个水平上的螺旋化（图 2.3）：

▶ **核小体**（nucleosome）是染色体组装的最基本单位。它由 8 个**组蛋白**（histone）组成核心，组蛋白是一种小的高度保守的 102～135 个氨基酸的基本（＝正电荷）蛋白质，每个核心包含 H2A，H2B，H3，H4 组蛋白的各两个分子。核心的周围是一段 146bp 的 DNA 双链绕其缠绕 1.75 圈，相邻的核小体之间由一段短距离的 DNA 相连，电子显微图像适合的标本显示 一‘串珠状’形象。

注：精细胞是不同的。精子细胞头部非常紧密组装的 DNA 是通过选择一类小的基本蛋白称为**鱼精蛋白**（protamine）来替换组蛋白而实现的。

▶ 直径 10nm 的串珠螺旋成直径 30nm 的**染色质纤维**（chromatin fiber）。间期染色体似乎就是由这些染色质纤维所组成的，并可能组成正如下述的长环。

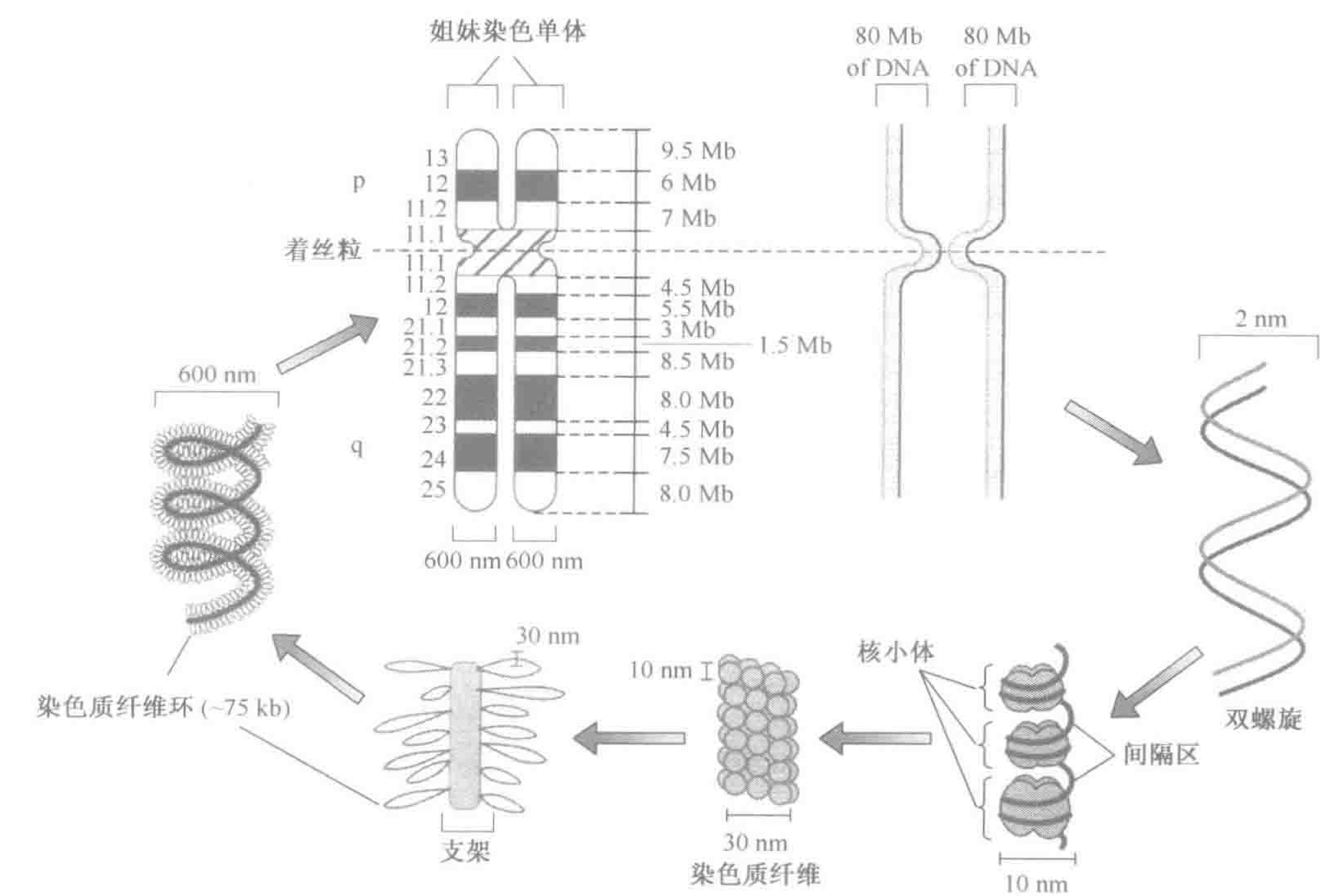


图 2.3　从 DNA 双螺旋到中期染色体

此图所示人类 17 号染色体 G-显带所见的 400 条带模式图。估计人类染色体的包装比例（线性 DNA 双螺旋的压缩程度）与核小体是 1∶6，与 30nm 纤维是 1∶36，与中期染色体是＞1∶10000。目前，中期染色体的着丝粒 DNA 的复制是否不同于染色单体的其他部分而延迟，或者是否全部 DNA 复制都发生在 S 期和着丝粒的缢痕由于某些其他原因还未确定。

在细胞分裂过程中，染色体变得更加高度浓缩。中期染色体的 DNA 凝集成它伸展时长度的大约 1/10 000。30nm 的染色质纤维环，每环含有 20～100kb 的 DNA，附着到中心的**支架结构**（scaffold）上，这是由非组蛋白酸性蛋白质组成。值得注意的是**拓扑异构酶Ⅱ**（topoisomeraseⅡ），该酶具有一种有趣的能力，通过切割一个 DNA 双螺旋使另一个双螺旋通过，并能修复它。已知拓扑异构酶Ⅱ和其他一些染色质蛋白质与富含 AT 序列相结合，而染色质环可由高度富含 AT（＞65％）的几百个碱基对的序列相附着（**支架附着区**，scaffold attachment region）。中期染色体的染色单体中，此环-支

架复合体通过螺旋化还可进一步紧缩（图 2.3）。

2.2.2　在间期核中单个染色体占有非重叠区

长久以来我们已知细胞质内具有高度的组织，但细胞核也有相当的亚结构。除了熟知的**核仁**（nucleolus，rRNA 在此转录并组装为核糖体亚单位）外，最近已鉴定了许多其他亚核成分（Cajal 体，PML 体，Paraspeckles 等，框 3.1）。在核内定位的染色体也是高度组织性的，并已开发出专门的技术来追踪间期活细胞内每个染色体的活动。

从某种程度来说，有丝分裂对间期染色体排列方向设置了限制。就在细胞分裂前，微管附于着丝粒并且牵引每条染色体朝向**有丝分裂纺锤体**（mitotic spindle）的两极之一移动（在有丝分裂期间定位染色体的微管网络——见框 2.1）。在染色体的移动中，着丝粒引导以染色体臂拖尾向的方式形成 V 形。在间期开始时，染色体倾向于保持这种所谓的**拉布尔定向**（Rabl orientation），着丝粒排列起来面向核的一极而端粒面向相反的一极（在间期短的一些细胞中，染色体倾向于在整个间期都采用这种方向）。

在哺乳动物中，间期细胞染色体一般看不见拉布尔定向，而且不同的着丝粒也不能很好地排成直线（虽然染色体在 S 期变得很分散之前，它们在 G_1 期的确是趋向簇集于核的周边）。虽然每条染色体处于高度伸展的形式，但染色体并无广泛的缠绕。相反，它们似乎有相对小的非重叠的**染色体区域**（chromosome territory）（图 2.4；Cremer and Cremer，2001；Parada and Misteli，2002）。

虽然间期染色体并未表现出在核内占据有利位置，不过染色体定位并不是随机的。例如，我们已知人类最富含基因的染色体集中于核的中心，而基因少的染色体则靠近核膜（图 2.4）（Boyle *et al.*，2001；Parada and Mistelli，2002）。染色体运动可能受限于与核膜和包括核仁在内的核内结构的相互作用（就染色体中含有核糖体 RNA 基因来说）。

2.2.3　染色体作为功能性细胞器：着丝粒的重要作用

正常染色体有一单个着丝粒，在显微镜下可以看到，称为**主缢痕**（primary constriction），姐妹染色单体在此区连接。着丝粒是细胞分裂过程中染色体分离所必需的，缺乏着丝粒的染色体片段（**无着丝粒片段**，acentric fragment）不会附着有丝分裂纺锤体（表 2.1 和图 2.7），所以不能进入至任何一个子细胞的核中。

框 2.1　有丝分裂纺锤体及其成分

有丝分裂纺锤体（mitotic spindle）（见下图）是由微管（由 α 和 β 微管蛋白重复异源二聚体构成的多聚体）和微管相关蛋白构成。两个纺锤体极的每一个都定义为**中心体**（centrosome），是主要的**微管组织中心**（microtubule organizing center）。中心体抽出向外生长的极性微管纤维，在中心体端为负（－）末端，而生长的远末端为正（＋）末端。每一个中心体都是由一个纤维性的基质（由大约 50 个拷贝的 γ-微管蛋白环复合体组成）构成的，在这个复合体中嵌埋着一对中心粒（见下图 A）。中心粒是由微管和相关蛋白组成的短圆柱状结构。此一对的两个中心粒总是彼此排列成直角而形成 L 形（下图 A）。在 G_1 期的一定时刻一对的两个中心粒分离，而在 S 期的子代中心粒开始在每个亲代中心粒的基础上以直角生长，直到 G_2 期完全形成。这两对中心粒仍以一单个的中心体复合体紧密在一起。直到 M 期开始，这个复合体分裂成两个，而其两半开始分离。此时每个中

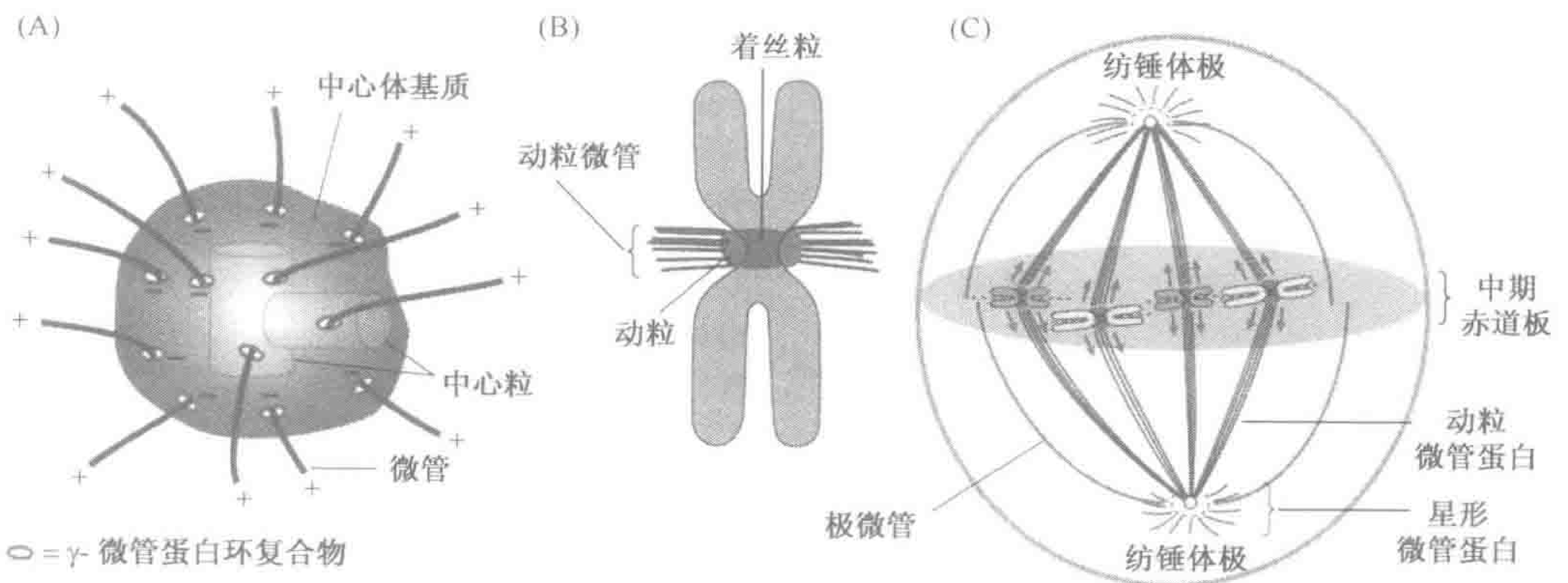
框 2.1　有丝分裂纺锤体及其成分（续）

心体发育自身的辐射阵列的微管（星体），同时开始向细胞相反的两端迁移并在那里形成**纺锤体极**（spindle pole）（下图 C）。纺锤体中有三种不同形式的微管纤维（下图 C）：

▶ **极纤维**（polar fiber）　从纺锤体的两极向赤道扩展。在前期，当核膜仍然完整时它们就开始发育。注：为了清晰的缘故，只显示了一对这样重叠的纤维。

▶ **动粒纤维**（kinetochore fiber）　直到前中期才发育。这些纤维附着在**动粒**（kinetochore）上，动粒是附着在每条染色单体着丝粒上的一个大的多蛋白结构（下图 B），并且在纺锤体极的方向扩展。

▶ **星体丝**（astral fiber）　围绕每一个中心体形成并向周边扩展。



(A)（中心体结构）、(B)（动粒-着丝粒联合）、(C)（有丝分裂纺锤体结构）

在有丝分裂晚前期，一对大的多蛋白质复合体称为**动粒**（kinetochore），在每一个着丝粒处形成。每一条姐妹染色单体附着一个，微管附着每一个动粒，连接每一条染色体的着丝粒和两个纺锤体极（框 2.1 和图 2.7）。在后期，动粒微管将两条姐妹染色单体拉向纺锤体相反的两极（图 2.7）。通过控制附着微管的组装、分解和运动分子的参与，最终驱动染色体运动，动粒在这一过程中发挥核心作用。

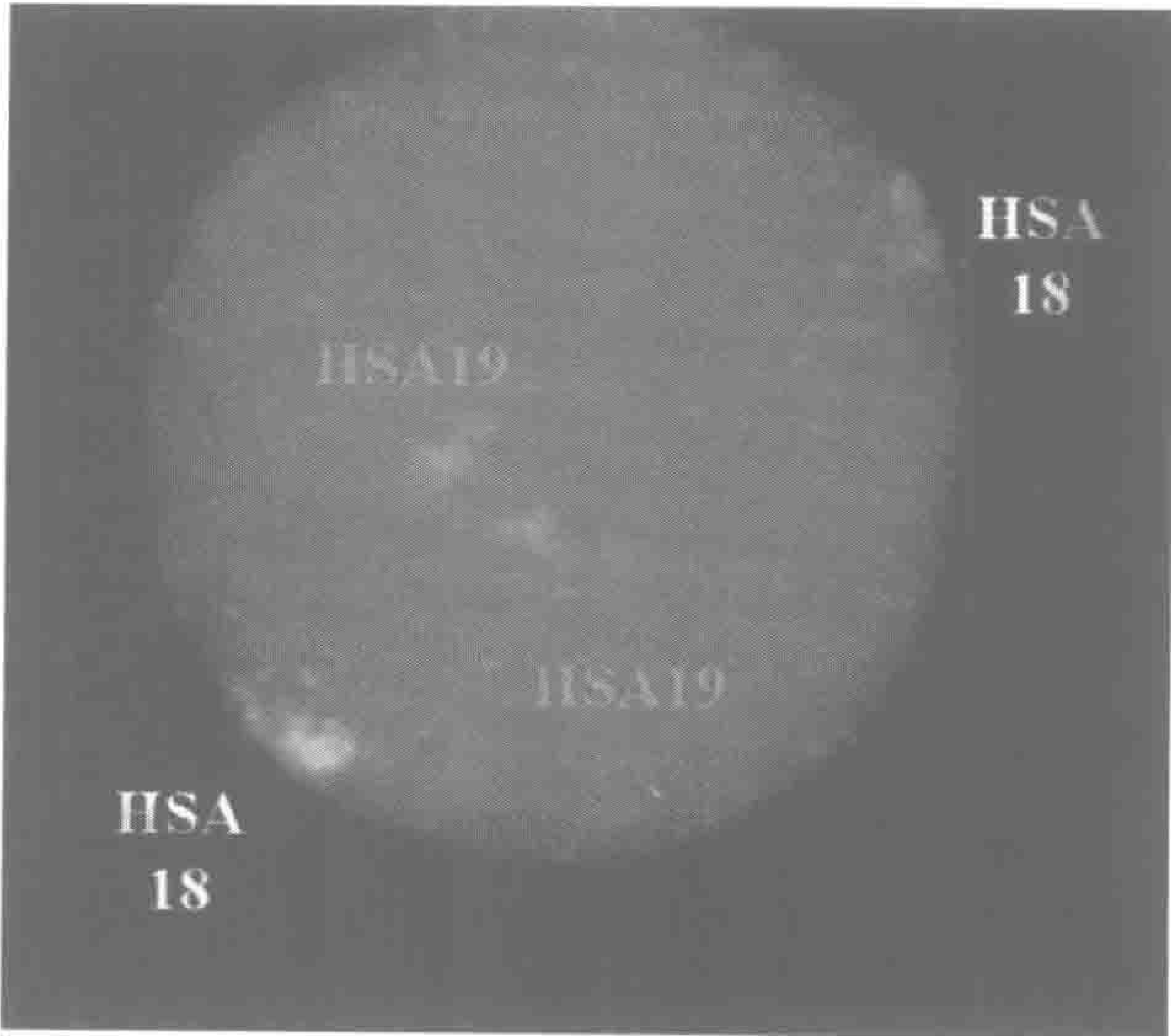


图 2.4　间期核中每个染色体占据不同的染色体区域
此图表示**染色体涂染**（chromosome painting）的例子（节 2.4.3）。这两条染色体在间期核中一条对 HSA18 是特异的（即人类 18 号染色体为绿色信号，位于核周边）而另一条对 HSA19 是特异的（人类 19 号染色体为红色信号，位于核内）。核用 DAPI 复染而显示蓝色（DAPI 是一种结合 DNA 的荧光染料）。图由 MRC Human Genetics Unit，Edinburgh 的 Wendy Bickmore 博士友好提供。

特异的 DNA 序列可能决定着丝粒的结构和功能。在简单的真核生物中，特化着丝粒功能的序列非常短。例如，在酵母（*Saccharomyces cerevisiae*）中的

着丝粒元件（CEN）大约长110bp，包含两个9bp和11bp长的高度保守旁侧翼元件和中心约80～90bp长的富含AT片段（图2.5）。这样细胞的着丝粒是可以互换的——来源于一条酵母染色体的CEN片段可以替换另一条染色体上的着丝粒而没有明显的后果。

在哺乳动物中，每个着丝粒的DNA由数百Kb的重复DNA组成，这些重复DNA，一些是染色体特异性的而另一些是非特异性的。人类着丝粒DNA的主要成分是**α-卫星DNA**（α-satellite DNA），基于一171bp的单体的串联重复DNA的复合家族（节9.4.1对人类着丝粒序列有更全面描述）。已知多种不同的蛋白质与人类着丝粒有关，包括直接与α-卫星DNA结合的CENP-B（表2.1）。令人惊奇的是，已知整个真核细胞中染色体分离装置是高度保守的，而着丝粒DNA和相关蛋白进化迅速，甚至被认为是负责产生物种的生殖隔离（Henikoff *et al.*,2001）。

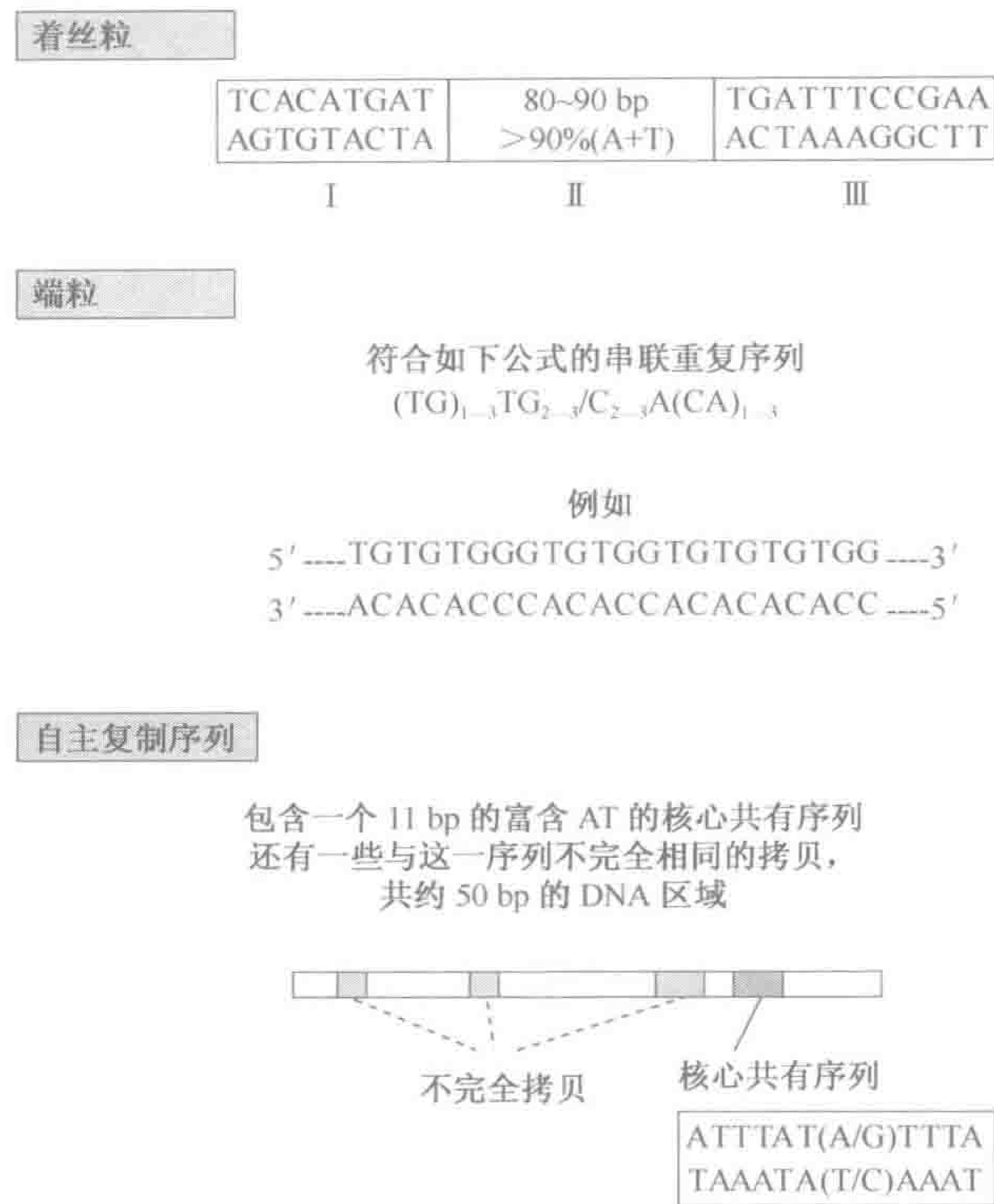


图2.5　酵母染色体的功能元件

表2.1　主要的人类着丝粒蛋白质

名称	位置	特性
CENP-A	动粒板外	着丝粒特异组蛋白H3变异体，对于激活CENP-C靶向动粒是必需的
CENP-B	着丝粒异染色质	结合在一个17bp序列α-卫星单体中（CENP-B框）；对于着丝粒异染色质的结构是必需的？
CENP-C	动粒板内	对于着丝粒或动粒的功能是必要的
CENP-G	动粒板内	与α-卫星家族序列的一个亚家族结合

注：CENP-E（在纤维冠中）、CENP-F（动粒板外）蛋白质和INCENP蛋白质（着丝粒异染色质）与着丝粒有短暂的联系。详情见Craig等（1999）.

2.2.4 染色体作为功能性细胞器：复制的起点

DNA 在大多数二倍体细胞的正常复制为每一细胞周期只复制一次。复制起始是由靠近 DNA 合成起点的顺式作用序列控制，这些可能是反式作用蛋白的结合位点。真核细胞的复制起点在酵母中研究得最详尽，推测的存在复制起点可通过遗传分析来检验。为了测试酵母 DNA 的随意片段启动自主复制的能力，将它同酵母细胞生长所必需的一个酵母基因一同整合到细菌质粒。通常以此构建（重组体）转化缺乏上述必需基因的突变型酵母。被转化的细胞只有在酵母细胞内的质粒能够复制时才能形成集落。然而，质粒中的细菌复制起点在酵母中不起作用。因此少数以高效转化的质粒必须具有一段插入的酵母片段的序列，来提供额外的染色体高效复制的能力，即所谓的**自主复制序列**（autonomously replicating sequence，ARS）元件。

ARS 元件被认为源于真正的复制起点。在某些情况下，通过将一个特异的 ARS 元件定位于特定的染色体位点并证实 DNA 复制的确在这个位置起始。酵母 ARS 元件长度约为 50bp，并由一个富含 AT 区所构成，该区含有保守核心一致序列及其一些不完整的拷贝（图 2.5）。此外，ARS 元件包含一个转录因子的结合位点，并且一个已知的多蛋白复合体也结合到这个起点。

由于缺乏遗传的检测方法，对于哺乳动物 DNA 的复制起点阐述的很少。一些起点曾经被研究，但是这些研究并不能鉴定单一的复制起点。这就导致了一种猜测，复制可以在几万个碱基对以上长度的区域内的多位点上起始（Gilbert，2001）。哺乳动物人工染色体似乎无需提供特异的 ARS 序列就可工作。

2.2.5 染色体作为功能性细胞器：端粒

端粒结构、功能和进化

端粒是一种由 DNA 和蛋白质组成的，为真核细胞染色体末端加帽的特殊结构。它们有几种可能的功能：

- **保持结构的完整性** 如果端粒丢失，会导致染色体末端的不稳定，它有与其他断裂染色体末端融合而涉及重组事件或被降解的趋势。端粒结合蛋白识别端粒突出的 3′端（见下文），在体外、也可能在体内保护末端的 DNA。
- **确保完全的 DNA 复制**（见以下关于端粒酶一节）
- **染色体定位** 端粒帮助建立细胞核的三维结构和/或染色体配对。在一些细胞，染色体末端似乎被拴系在核膜上，提示端粒可以帮助定位染色体。

真核细胞的端粒由一简单中等长度的串联重复序列构成，在该序列的一条 DNA 链上富含 TG，而在互补链富含 CA。在人类（和其他动物），此重复序列是六核苷酸 TTAGGG，人类端粒的（TTAGGG）$_n$ 序列典型的长度约为 3～20kb。除端粒外，在与任何单一序列相遇之前（在靠近着丝粒的方向）有一100～300kb 的**端粒相关重复**（telomere-associated repeat）序列。

与着丝粒不同，端粒序列在进化上是高度保守的，不同物种端粒中的简单重复序列是非常相似的。例如，在草履虫（*Paramecium*）中是 TTGGGG，在锥虫（*Trypanosoma*）中是 TAGGG，拟南芥（*Arabidopsis*）中是 TTTAGGG。而在人类中

是 TTAGGG（注：然而，在一些物种中精确重复单位有一些可塑性，就像在酿酒酵母中的情况，图 2.5）。此外，在各种类型的端粒结合蛋白中也有保守结构（Blackburn，2001）。然而，真核生物中的端粒相关重复并不保守而它们的功能也尚不知晓。

端粒酶和末端复制问题

在 DNA 复制期间，后随链是分段合成的，这是因为它必须沿着与 DNA 合成的 5′→3′的方向相反的方向生长，一连串的“后期缝合”（back-stitching）合成需要产生一系列的 DNA 片段，然后它们的末端被 DNA 连接酶所封接（图 1.9 和节 1.2.2，节 1.2.3）。与 RNA 聚合酶不同，DNA 聚合酶绝对需要双链核酸的游离 3′羟基，并由此延伸合成。要达到这一点，可用一个 RNA 聚合酶来合成一互补的 RNA 引物，以此首先合成每一个用于后随链合成的 DNA 片段。在这些情况下，RNA 引物需要在拷贝此序列之前存在一些 DNA 以作为它的模板。然而，在线性 DNA 分子的最末端绝不能是这样的模板，这就需要有不同的机制来解决线性 DNA 分子末端的复制问题。

末端复制（end-replication）问题已经通过使用一种特殊类型的反转录酶（RNA 依

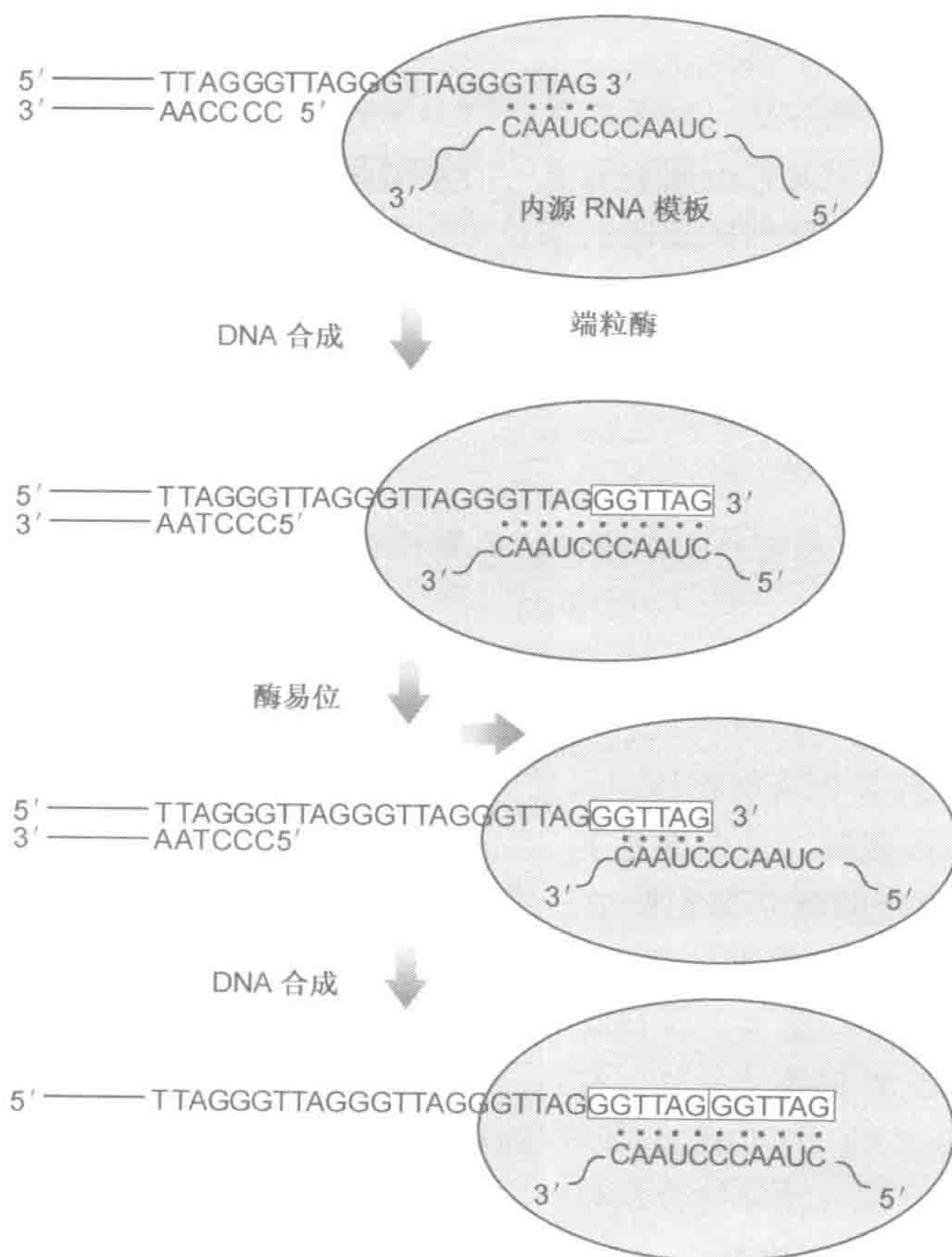


图 2.6 端粒酶利用内源性 RNA 模板合成 DNA 来延长富含 TG 链的端粒

注：暗影显示新合成的六核苷酸以及延伸机制是依赖于含有一个几乎完全串联重复的 RNA 模板，如下列序列中下划线所示：5′CUAACC CUAAC 3′。

赖的DNA聚合酶）延伸前导链的合成而得到解决，该酶是由一种特殊的RNA-蛋白质酶称为**端粒酶**（telomerase）所提供的。端粒酶在近其5′端携有它的RNA成分的一短序列CUAACCCUAAC，内有一六核苷酸序列（折叠型），它是人类端粒重复序列（TTAGGG）的反义序列。这一序列将作为模板在前导链上起始端粒DNA序列的DNA合成，前导链的进一步延伸为DNA聚合酶α完成后随链的合成提供了必要的模板（图2.6）。这一机制使得端粒自身留下了一段提供单链DNA靶点突出的3′端，与端粒特异蛋白诸如人类TRF2相结合。然而，端粒序列的实际性质可能并不重要。已知端粒的长度是高度可变的并从属于遗传的控制（节17.5.1）

2.2.6　异染色质和常染色质

在间期的核内，大部分染色质以伸展状态存在，分散在整个细胞核并且着色弥散（**常染色质**，euchromatin）。然而，一些染色质在整个细胞周期中都保持高度浓缩，形成深染区域（**异染色质**，heterochromatin）。位于常染色质上的基因可能表达也可能不表达，这取决于细胞类型及其代谢的需要；但是位于异染色质内的基因或是自然地或是染色体重排的结果，很可能是不表达的。异染色质有两类：

- **组成性异染色质**（constitutive heterochromatin）通常是无活性和凝缩的。它由大量的重复DNA组成，位于染色体着丝粒内及其周围和其他一定的区域（见图2.15组成性异染色质在人类染色体上的位置和表9.2有关DNA量的估计）。
- **兼性异染色质**（facultative heterochromatin）既可以活性形式（去浓缩的）、也可以失活形式（浓缩的）存在。包括哺乳动物X染色体的失活（节10.5.6）或在雄性减数分裂期间性染色体的沉默的例子。后者的情况，雄性减数分裂期间X染色体和Y染色体两者都是失活的（在人类大约是15天的一段期间），它们变得浓缩而形成XY小体，被分离到一个特殊的核区域。

在常染色质中，**G带**（G band）显示了一些异染色质的特性，但是程度少一些（G带是由于Giemsa染料阳性着色而呈现的染色体深带，浅染的Giemsa阴性带则可当作R带—节2.4.1）。在中期染色体，G带染色质比R带染色质更加浓缩，且G带上的基因相当贫乏，而由T显带揭示在R带的亚组上基因的密度特别高。节1.3.5讨论了转录激活和失活的染色体区域染色质的不同结构。

2.3　有丝分裂和减数分裂是细胞分裂的两种类型

2.3.1　有丝分裂是细胞分裂的正常形式

当个体的发育从胚胎开始经过胎儿、幼儿、到成年需要通过细胞分裂产生所需的大量细胞。此外，许多细胞具有有限的生命周期，因此成年人需要持续不断地产生新细胞。所有的这些细胞都是通过有丝分裂而产生的。从受精卵卵裂到一个人的死亡，有丝分裂是细胞分裂的正常过程。在人的一生中，可约有10^{17}次有丝分裂（节11.2.1）。

细胞周期的M期（图2.1）由核分裂的不同阶段（有丝分裂的前期、前中期、中期、后期、末期）和与有丝分裂的最终阶段相重叠的细胞分裂（**胞质分裂**，cytokinesis）阶段组成（图2.7）。在细胞分裂的准备阶段，先前高度伸展的染色体收缩、浓缩，

以至到有丝分裂的**中期**（metaphase），在显微镜下它们可以很容易看到，即使 DNA 在之前的一段时间已经复制了，也只有在前中期才能看到附着于着丝粒的两条**姐妹染色单体**（sister chromatid）组成的每个染色体。

在不同纺锤体纤维之间的相互作用下（框 2.1），将染色体拉向中心，通过有丝分裂的中期染色体成一行排列在赤道面（**中期板**，metaphase plate）上。注：在有丝分裂期间二倍体组中的每一条染色体独立行动，父体和母体的同源染色体毫不相关。在后期，着丝粒分裂导致先前的姐妹染色单体的物理分离，并由纺锤丝牵引以确保分离的姐妹染色单体移向相反的两极（图 2.7 和图 2.8）。除非在 DNA 复制时发生任一错误，否则两条姐妹染色单体的 DNA 是相同的。因此有丝分裂的作用就是产生含有精确相同组成的 DNA 序列的子细胞。

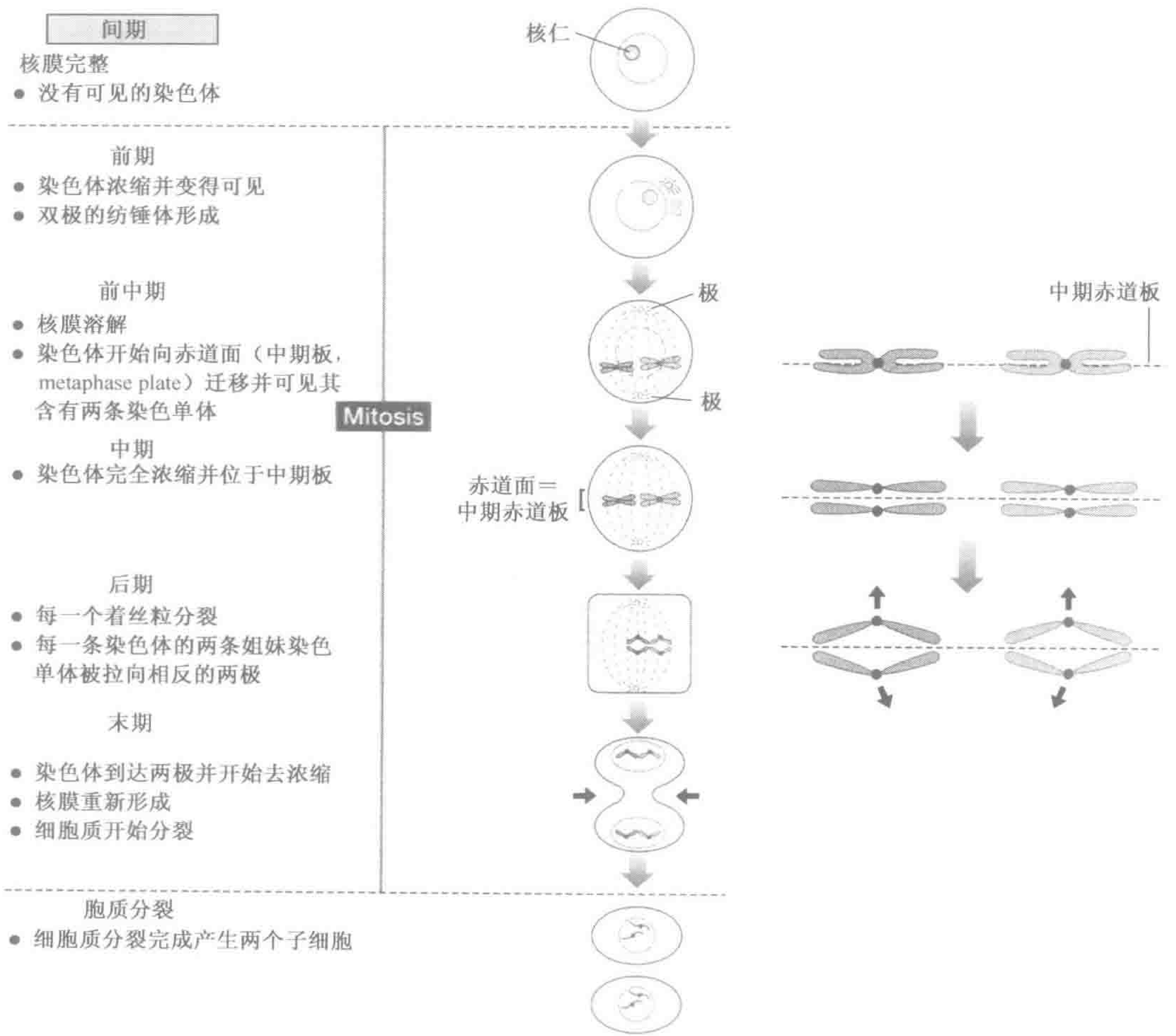


图 2.7 细胞分裂（有丝分裂）略图

2.3.2 减数分裂是产生精子和卵细胞的一种特殊的细胞分裂形式

原始生殖细胞迁移入胚胎生殖腺内并且进行连续重复循环的有丝分裂，在女性形成

卵原细胞（oogonia），在男性形成**精原细胞**（spermatogonia，注：这涉及在男性比在女性有更多次的有丝分裂—框 11.4—这也许是解释突变率性别差异的一个重要的因素）。进一步的生长和分化，在卵巢产生**初级卵母细胞**（primary oocyte），在睾丸产生**初级精母细胞**（primary spermatocyte）。这些特化的二倍体细胞进行减数分裂（图 2.9）。

减数分裂涉及两次连续的细胞分裂，但 DNA 只复制一次，故其产物是单倍体。在男性产物是四个精子；而在女性存在**不对称细胞分裂**（asymmetric cell division）是因为每一个阶段细胞质的不均衡分裂：减数分裂Ⅰ（第一次减数分裂）的产物是一个大的**次级卵母细胞**（secondary oocyte）和一个小细胞（**极体**，polar body）。在减数分裂Ⅱ（第二次减数分裂）期间，次级卵母细胞产生一个大的成熟卵细胞和一个第二极体。在有丝分裂和减数分裂之间有两点重要的不同（表 2.2）：

- ▶ 有丝分裂的产物是二倍体，减数分裂的产物是单倍体；
- ▶ 有丝分裂的产物在遗传上是相同的，减数分裂的产物在遗传上则是不同的。

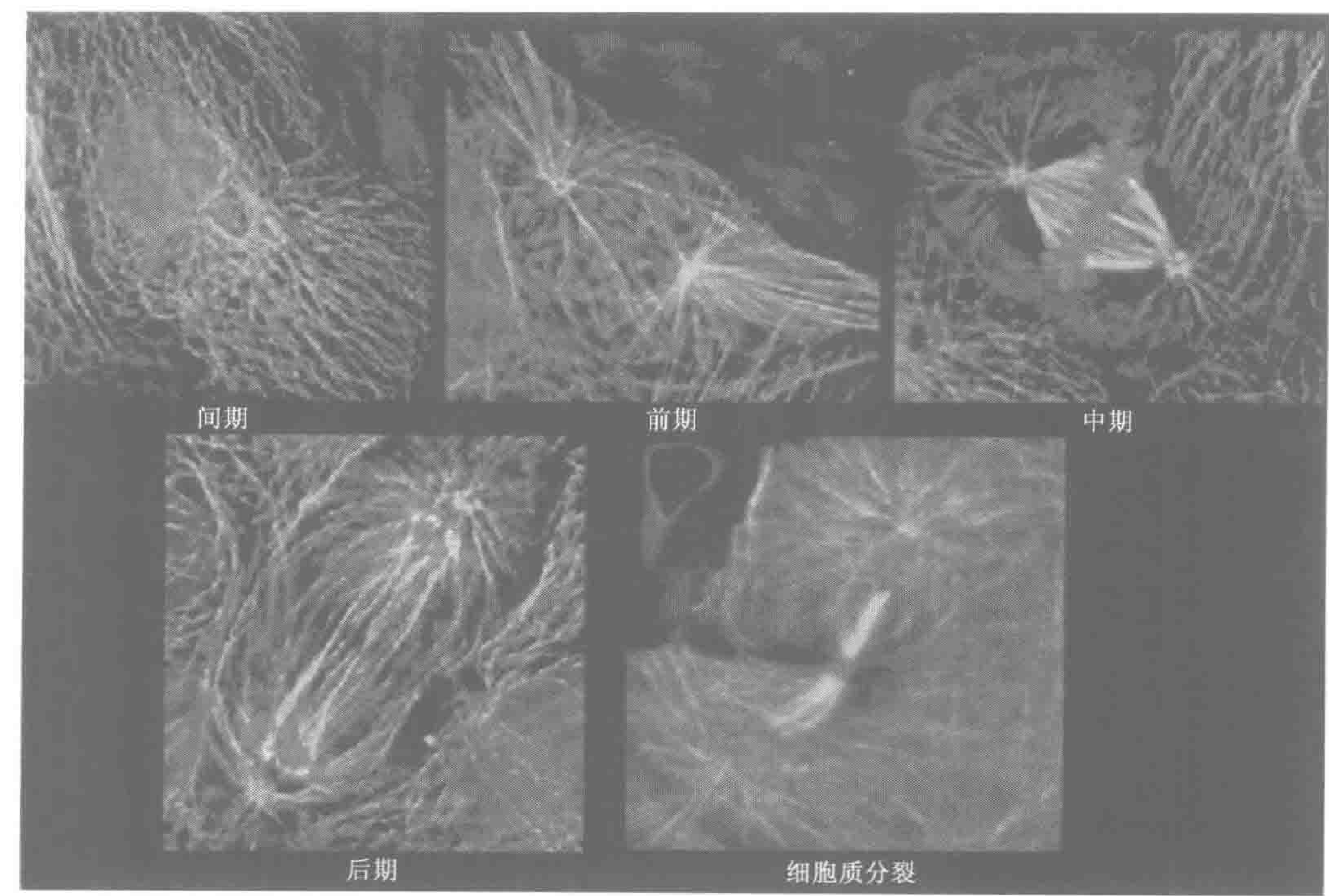


图 2.8　图 2.7 中细胞的生物学家所见

是用反卷积显微镜（deconvolution microscopy）获得的 Hela 细胞的图像，DNA 用 DAPI 染色（假染为红色），而微管用β-微管蛋白抗体染色（假染为绿色）。本图由 Edinburgh 大学 William Earnshaw 提供，经 Elsevier 允许，从 Pollard 和 Earnshaw《细胞生物学》(2002) 上复印。

有丝分裂涉及细胞周期的单一循环（图 2.1）。DNA 在 S 期中复制，在 M 期这两个拷贝在两个子细胞间极其相等。对于减数分裂，在分裂之前也进行一次 DNA 合成，但是随后就进行了两次细胞分裂，而在此之间没有 DNA 合成，以致产物最终是单倍体。第二次减数分裂与有丝分裂相同，而第一次分裂则有重要的不同，即其目的就是在

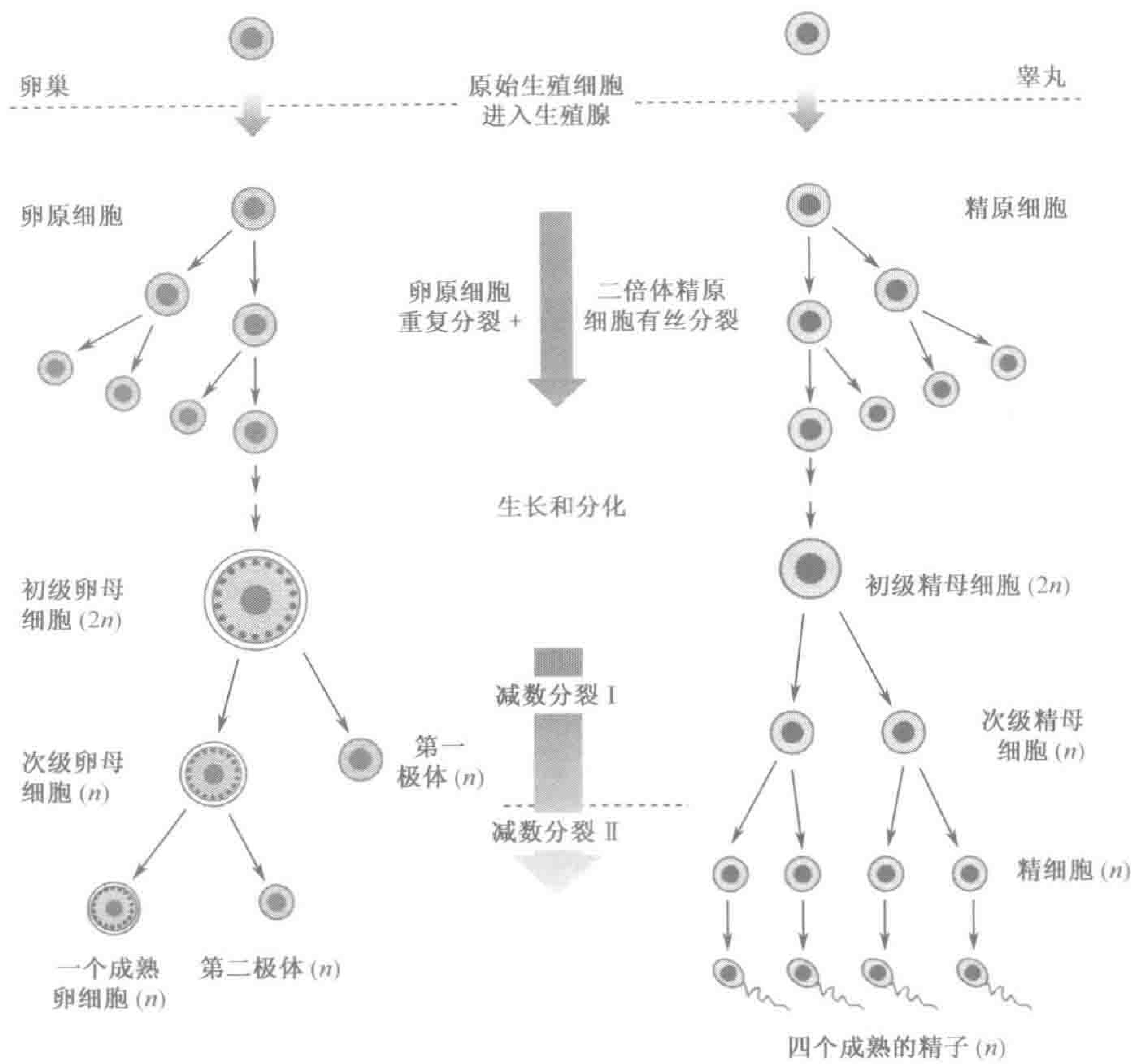


图 2.9 生殖系的发生

种系经过二倍体细胞的重复有丝分裂而发育，大量的产生初级卵母细胞和初级精母细胞。这些二倍体细胞可以进行减数分裂。减数分裂涉及两次细胞分裂但只有一次 DNA 复制期，所以产物是单倍体。在人类，初级卵母细胞在胎儿时期即进入减数分裂 I，但稍后便停滞在前期，一直到青春期或更晚。在这期间，初级卵母细胞完成它们的生长期，获得外层卵胶（膜）包被、皮质颗粒、核糖体、mRNA、卵黄等。在青春期之后，一个卵母细胞每个月完成减数分裂，而精子则从青春期连续不断的产生。

表 2.2 有丝分裂和减数分裂的比较

	有丝分裂	减数分裂
位置	所有的组织	只在睾丸和卵巢
产物	二倍体体细胞	单倍体精子和卵细胞
DNA 复制与细胞分裂	每次分裂正常复制一次	DNA 只复制一次但细胞分裂两次
前期的长度	短（在人类细胞～30 分钟）	减数分裂 I 长而复杂，费时数多年完成
同源染色体配对	无	有（在减数分裂 I）
重组	罕见且属异常	正常，每条染色体臂至少一次
子细胞间的关系	遗传上相同	不同（重组和同源染色体的）

子细胞之间产生遗传的多样性。这一点是通过两种机制完成的：父体和母体同源染色体的自由组合和重组。

父体和母体同源染色体的自由组合

在减数分裂Ⅰ期间，父体和母体每一对同源染色体配对形成一个**二价体**（bivalent）（**联会**，synapsis：图 2.11）。在 DNA 复制之后，每一条染色体由两条姐妹染色单体构成，所以二价体在中期板上是一四链结构。然后纺锤体纤维牵引着一个完整的染色体（两条染色单体）移向两极。然而，对于 23 对同源染色体每一条同源染色体选择哪一个子细胞是自由的。因此由一个人产生的亲体染色体可有 2^{23} 或大约 8.4×10^6 可能的组合（图 2.10）。

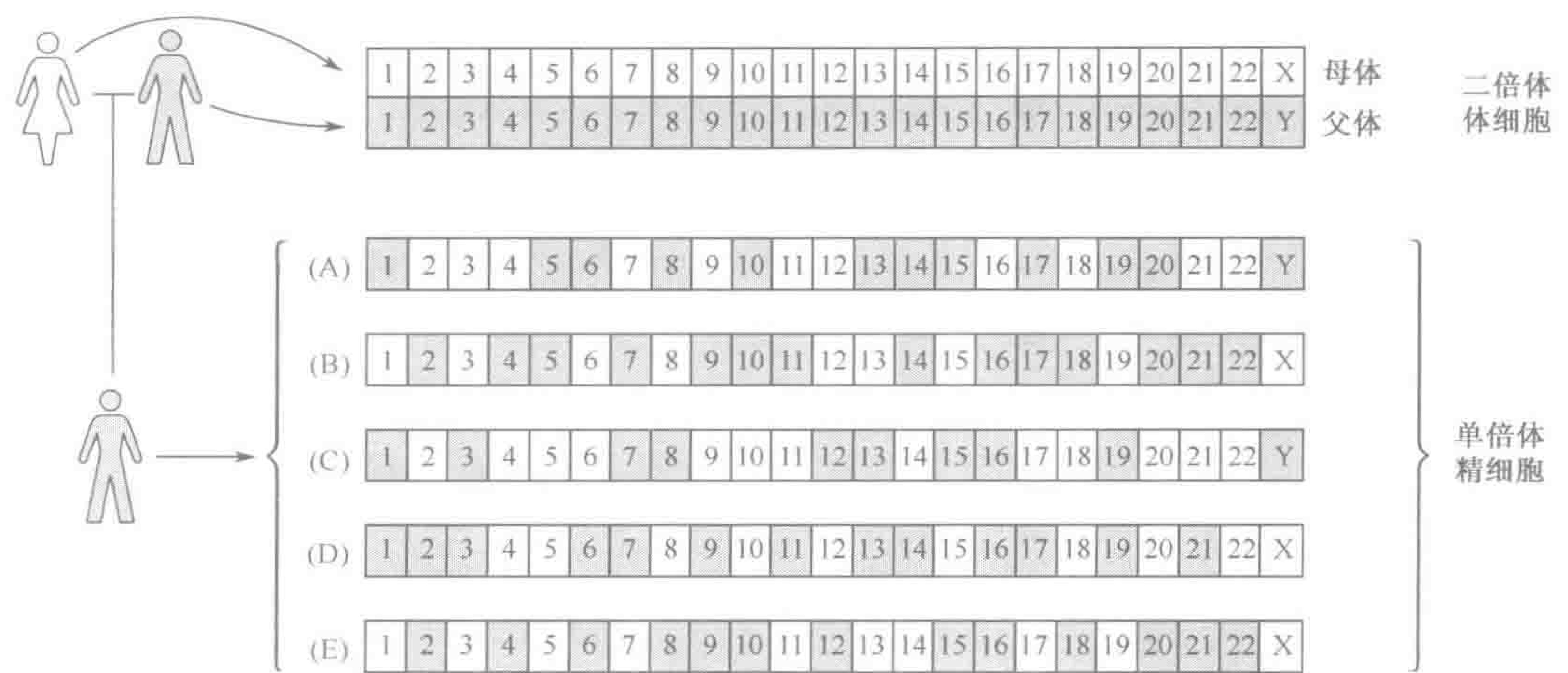


图 2.10　减数分裂：减数分裂Ⅰ期的父体和母体同源染色体的自由组合是产生第一水平的遗传多样性

从二倍体细胞的 23 对染色体中的任取一对染色体都可有 2^{23} 或 8.4 百万种组合。A 到 E 这五种配子正表明来自父体和母体染色体的五种可能的结合，此图解未考虑重组，而重组则是通过父体和母体序列的混合确保每条染色体的传递，是第二水平的遗传多样性。

重组

在减数分裂Ⅰ的前期，在每一个二价体中联会的同源染色体以随意的方式交换片段。在偶线期，每一对同源染色体开始由两条紧密附着的染色体形成**联会复合体**（synaptonemal complex），被一个长的线性蛋白核心所分隔。这个复合体的完成标志着粗线期的开始，而正是在这一阶段发生重组（交换）。交换涉及一条父体和一条母体染色单体双螺旋的断裂及其末端的连接。总之，在第一次减数分裂前期Ⅰ的同源染色体之间重组结合加上第一次减数分裂后期Ⅰ同源染色体的自由组合，确保了单一个体可以产生几乎无限的遗传上不同的配子。

同源染色体的线性排列的机制尚未阐明，然而，这种紧密并列方式被认为是重组所

必需的。**重组结**（recombination nodule）：位于联会复合物的多蛋白复合体，一个很大的多蛋白集合体位于联会复合体之间，并认为中介调节重组事件。这两条同源染色体在物理上能看到在特异位点连接，每一种连接被称之为**交叉**（chiama）（chiasmata，复数），并且标明一个交换点。男性在减数分裂的过程中每个细胞平均发生 55 次交叉，而女性减数分裂中可能是多于 50%。交叉的遗传学意义在 13 章叙述。

交叉除了在重组起的作用之外，还被认为是减数分裂Ⅰ期染色体正确分离的必要条件。将父体和母体同源的每对染色体维持在纺锤体上直至后期Ⅰ（图 2.11 和图 2.12），交叉具有与有丝分裂和减数分裂Ⅱ中着丝粒的相同作用。有遗传证据表明配子缺少二价体的交换常导致儿童染色体数目的异常。

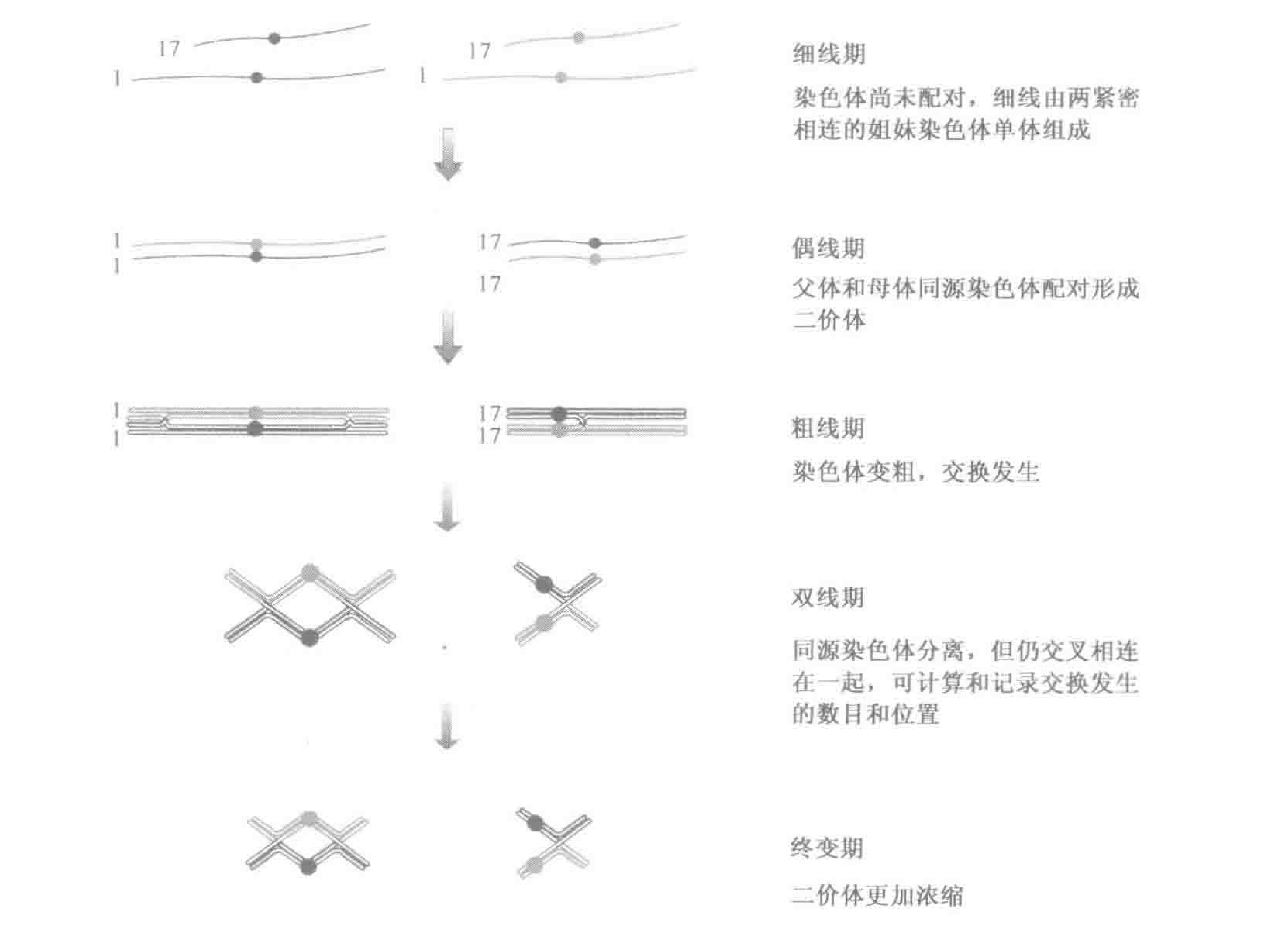


图 2.11　减数分裂：减数分裂前期Ⅰ的五个阶段

图示两条具代表性的同源染色体对。1 号染色体的二价体发生了 2 次交换而 17 号染色体的二价体发生了一次交换。为了清晰的辨认，1 号染色体的两次交换均涉及两个相同的染色单体。实际上，交换的次数可能更高，多次交换可能涉及二价体中的三条甚至四条全部染色单体，如图 13.2 所示。

减数分裂Ⅱ看似与有丝分裂相同，只是染色体数目为 23 条而不是 46 条。它每条染色体都由 2 条染色单体组成，并在后期Ⅱ分离。然而有一个区别，即有丝分裂染色体的姐妹染色单体是相同的，因为是相互拷贝的，但减数分裂Ⅱ的染色体的两条姐妹染色单体由于减数分裂Ⅰ交换的结果在遗传上可能是不同的（图 2.12）。

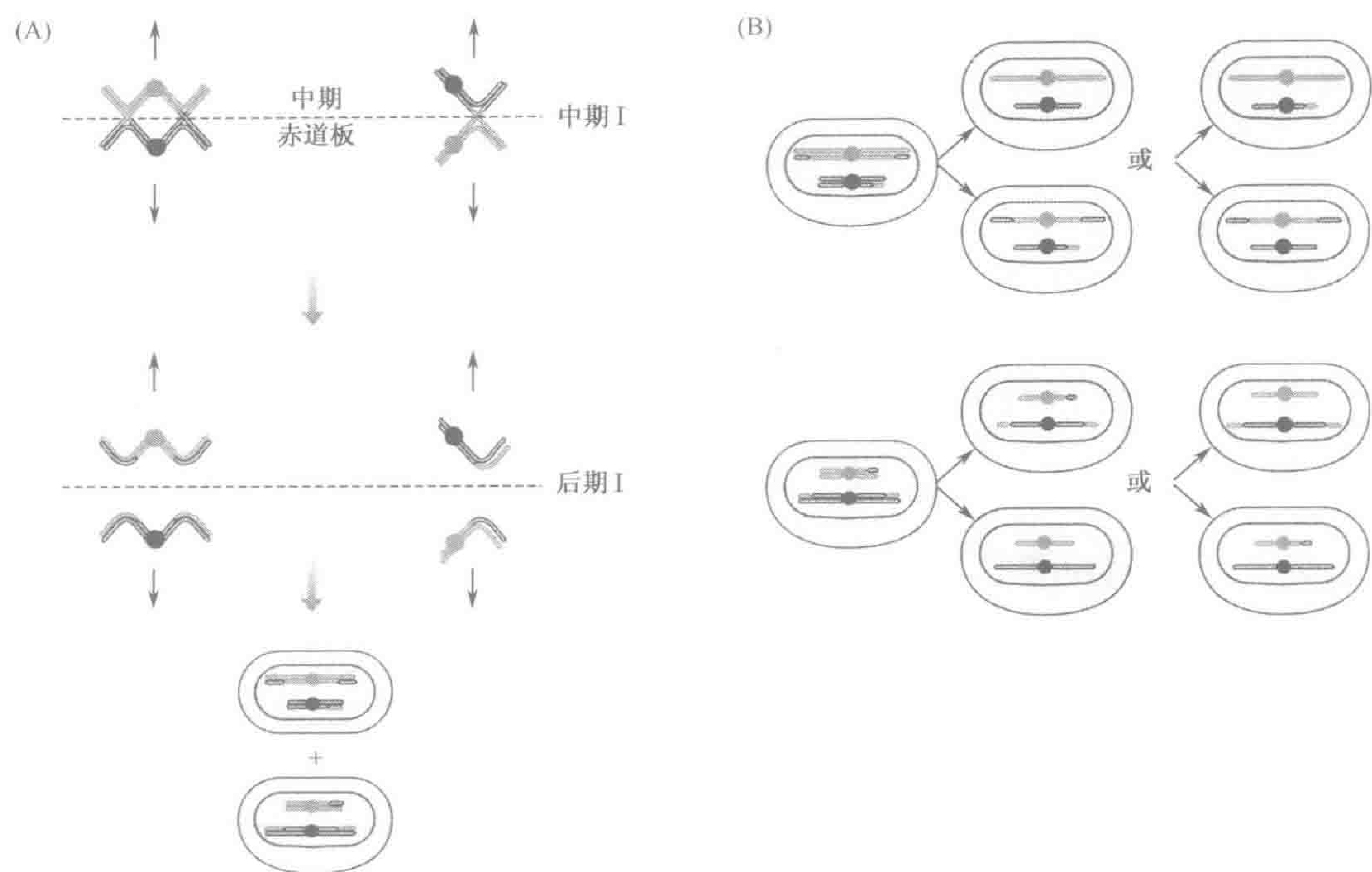


图 2.12　减数分裂：中期 I 至配子

此图接图 2.11

（A）从减数分裂 I 中期至细胞分裂，此图展示两个二价体一种可能的分离模式

（B）减数分裂 II，尽管每条染色体的两个姐妹染色单体将分离到不同的子细胞，但正如图中所示，不同染色体行为是独立导致许多重组是可能的。

2.3.3　X-Y 配对和假常染色体区域

在女性减数分裂中，每个染色体都有一个完全的同源伙伴，且两条 X 染色体与其他同源染色体一样进行联会和交换。在男性减数分裂中则存在一个问题，人类 X 和 Y 两条性染色体彼此有很大的区别。尽管如此，男性在前期 I 中，X 和 Y 染色体的确是成对出现的。因此可确保在后期 I 每个子细胞将接受一条性染色体——X 染色体或是 Y 染色体。X 和 Y 染色体的配对是末端对末端的，而不是沿着染色体的全长进行的，正因如此，在 X、Y 染色体短臂的末梢之间可能形成 2.6 Mb 的同源配对。配对受到在此区域的专性交换得以证实。在配对片段的基因有如下有趣的性质：

- 它们在 X 和 Y 染色体上以同源的拷贝存在；
- 它们不受 X 染色体失活机制的支配（不出所料，因为每条性染色体都有两个拷贝）；
- 由于交换的缘故，在这些座位的等位基因并不表现出正常的 X 连锁或 Y 连锁的遗传模式，但与常染色体上的等位基因一样分离。

由于这一特性，此区域被称为**主要假常染色体区域**（major pseudoautosomal region)。第二个 320kb 较小的假常染色体区域位于两个染色体的长臂的末梢，但在这个小的假性常染色体区域的配对和交换并无男性减数分裂专性特点。

2.4 人类染色体研究

2.4.1 在任一分裂细胞中均可见有丝分裂的染色体，但难以研究人类减数分裂的染色体

为染色体分析获取与制备人体细胞

正常情况下在分裂的细胞中可看到染色体，但难以直接从人体获取分裂中的细胞。骨髓是一个可能的获取来源，但它是更易采用到处可获取来源方便的未分裂细胞，并在实验室中进行培养。最常用于细胞遗传学分析的人类细胞来源就是血细胞和皮肤成纤维细胞。大多数人并不介意捐献几毫升的血液，而且用凝集素诸如植物血凝素容易诱导血液中的 T 淋巴细胞分裂。皮肤成纤维细胞经皮肤活检获取并培养。此外，常规的产前诊断涉及染色体分析可利用脱落羊水中的胎儿细胞或绒毛细胞。

虽然早在 19 世纪 80 年代，人们已经对一些有机体的染色体作了准确地描述，但是数十年来，所有试图制备分散的人类染色体产生无法分析的混乱。为获得可易分析的分散染色体的关键是一项新技术：即在液体中悬浮生长细胞，用低渗的盐溶液处理使细胞膨胀。这使得在 1956 年获得了第一个质量好的标本。将血液中的白细胞置于有植物血凝素的培养基中进行培养，使之生长 48～72h，在这期间细胞应该自由地分裂。尽管如此，因为 M 期只占细胞周期中的一小部分，所以实际上在任何时候细胞是分裂的很少。

有丝分裂率（mitotic index，有丝分裂时细胞的比率)，用破坏纺锤体的试剂诸如**秋水仙素**（colcemid）处理培养基，可提高有丝分裂率。处理后的细胞到达细胞周期的 M 期并停留在 M 期，所以细胞积累在有丝分裂的中期。通常情况下，人们更倾向于研究**前中期的染色体**（prometaphase chromosome)，此期的染色体收缩较少所以可显示更多的细节。通过暂时性的阻止细胞周期来使细胞生长同步化。通常这可借助加入过量的胸腺嘧啶核苷（它可以使 dCTP 减少而引起 DNA 合成减慢，使细胞停留在 S 期）来实现。当胸腺嘧啶核苷效应除去后，细胞便通过同步化周期而进展。人们通过实践和错误，可以确定去除效应后的一个时间，此时所希望的前中期阶段细胞可具有理想的比率。

人们只能在睾丸和卵巢的样本中研究减数分裂。由于仅在胎儿卵巢中才有活跃的减数分裂，所以对女性减数分裂的研究尤其困难。反之，可通过对任一青春期后男性志愿提供睾丸活检来研究男性精子的减数分裂。尽管这一研究的方法学很麻烦，但可通过对精子的染色体分析来研究减数分裂的结果。同时，减数分裂的分析可用于男性不育的某些研究。

核型与染色体显带

直至 20 世纪 70 年代，根据染色体的大小及着丝粒的位置才鉴定了染色体，这只能将染色体分成几组（表 2.3)，却不能进行明确的鉴别。染色体显带技术（框 2.2）的采用终于可以鉴别每条染色体，以及可以更精细地确定易位断裂点和亚染色体缺失等。通过运用更加伸长的染色体，例如前中期或是更早期（非中期）的染色体，可以提高染色体显带的分辨率。典型的人类染色体高分辨显带分析可以显示 400，550 或 850 条带（图 2.13 和图 2.14)。

表 2.3　人类染色体组

组	染色体	描述
A	1-3	最大；1、3 为中央着丝粒，2 为亚中着丝粒
B	4、5	大；亚中着丝粒，长短臂大小差别明显
C	6-12、X	中等；亚中着丝粒
D	13-15	中等；近端着丝粒，有随体
E	16-18	小；16 为中央着丝粒，17 和 18 为亚中着丝粒
F	19、20	小；中央着丝粒
G	21、22、Y	最小；近端着丝粒 21 和 22 有随体，而 Y 没有

除了 21 号染色体小于 22 号染色体外，人类常染色体均从最大到小来编号。

框 2.2　染色体显带技术

G 显带（G-banding）：染色体用胰蛋白酶来消化控制，然后用 Giemsa（一种 DNA 化学染色剂）染色。阳性深染的称为 G 显带，浅染的为 G 阴性带。

Q 显带（Q-banding）：使用可以优先结合富含 AT 的 DNA 的荧光染料诸如喹吖因、DAPI（苯茚二酮）或 Hoechst33258 对染色体进行染色，然后用紫外线荧光显微镜进行观察。有荧光的带称为 Q 带，其标记相同的染色体节段为 G 显带。

R 显带（R-banding）：它主要与 G 显带模式相反，用 Giemsa 染色之前先将染色体加热变性。热处理富含 AT 的 DNA 其 R 带为 Q 阴性。用 GC 特异性染色剂诸如色霉素 A3、橄榄霉素或金霉酸染色可产生同样的模式。

T 显带（T-banding）：用来鉴别那些特异集中分布在端粒区的 R 带的亚组。T 带是 R 带中最深染的那部分。可在 Giemsa 染色之前进行特别的热处理或者是用染色剂和荧光剂组合来进行观察。

C 显带（C-banding）：为了证实着丝粒为组成性异染色质的一种显带技术。在 Giemsa 染色之前要用氢氧化钡溶液处理使 DNA 变性。

对于染色体的组成可以用**核型**（karyotype）来描述，核型说明了染色体的总数和性染色体组成。人类正常男性和女性的核型分别是 46，XY 和 46，XX。当出现一条染色体畸变，核型描述异常的类型和受累染色体带及亚带。细节见框 2.3 的**染色体命名**（chromosome nomenclature）。

染色体用**核型图**（karyogram）来显示（图 2.14，通常不严谨的称为核型）。以前，通过剪切分散的染色体的图片配对同源染色体。现在，使用带有图像分析程序的计算机来制作核型图。

如框 2.2 所示，**染色体显带**（chromosome banding）需要将染色体进行变性和/或酶消化处理，继而用 DNA-特异的染料染色：结果使得人类或其他有丝分裂染色体染成一系列明暗相间的条带（图 2.13 及图 2.14；Craig and Bickmore，1993）。显带带型可为 1～10Mb 以上区域结构的一些性质提供证据，故而是有意义的（同样对细胞遗传学也是有用的）。显带带型也与其他的性质相关，深染的 G 带区域在细胞周期的 S 期复制较迟且含有较多的浓缩的染色质，而 R 带（=浅 G 带），在 S 期一般复制较早且含有较少的浓缩的染色质。基因大部分集中在 R 带，而迟复制、多浓缩的 G 带的 DNA 则少有转录活性，此外 G 和 R 带分散的重复元件类型也不同（节 9.5.2，节 9.5.3）。

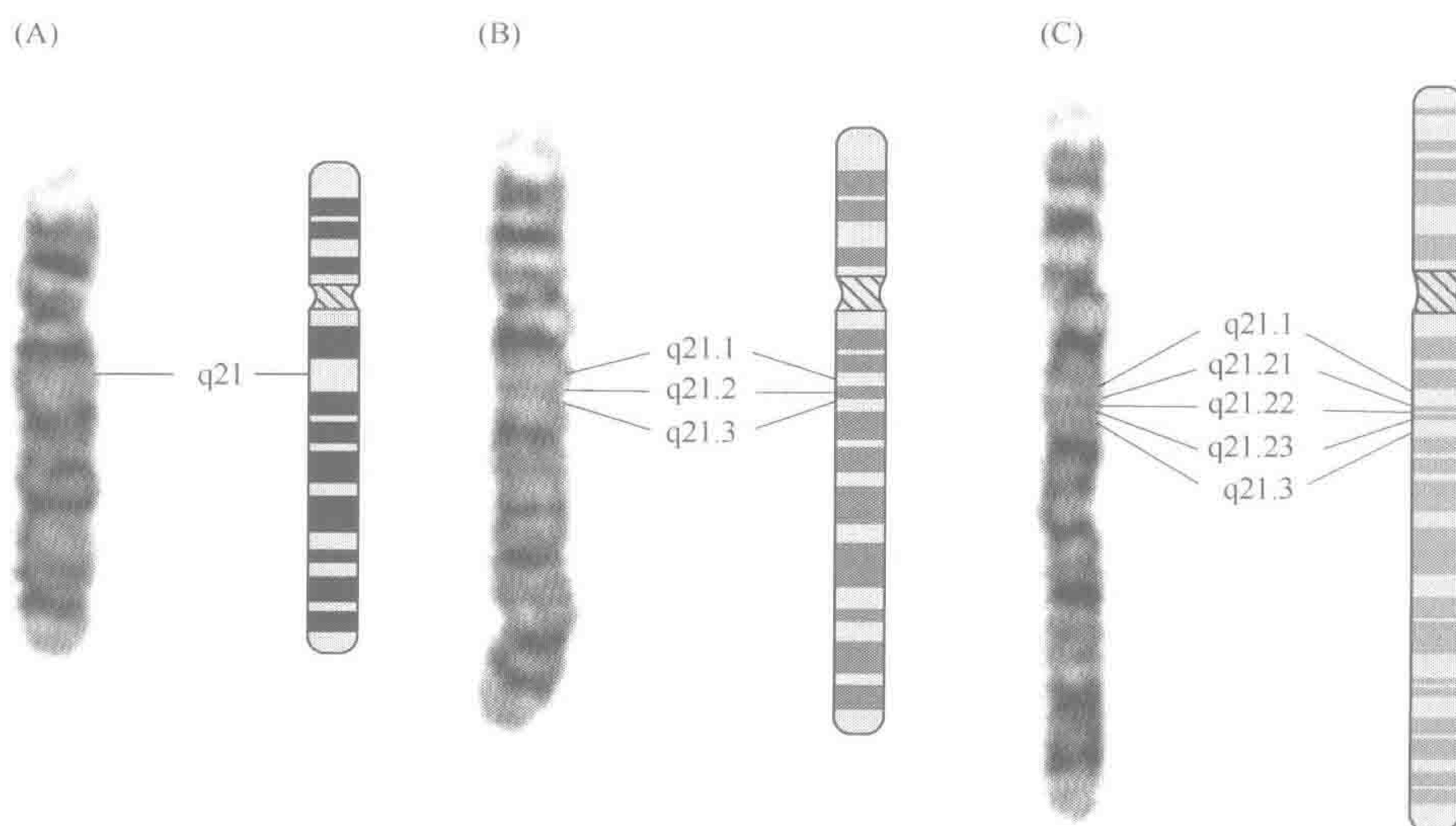


图 2.13　不同分辨率的染色体显带可分辨带、亚带及次亚带

4 号染色体（与其相伴的模式图）所示，随着分辨率水平的提高，每个单倍体组大约（A）400，（B）550 和（C）850 带。**注**：当分辨率提高时带分为亚带。选自 Cross 和 Wolstenholme（2001）。引自人类细胞遗传学：结构分析，3rd Edn（ed. D. E. Rooney），经 Oxford University Press 同意而复制。

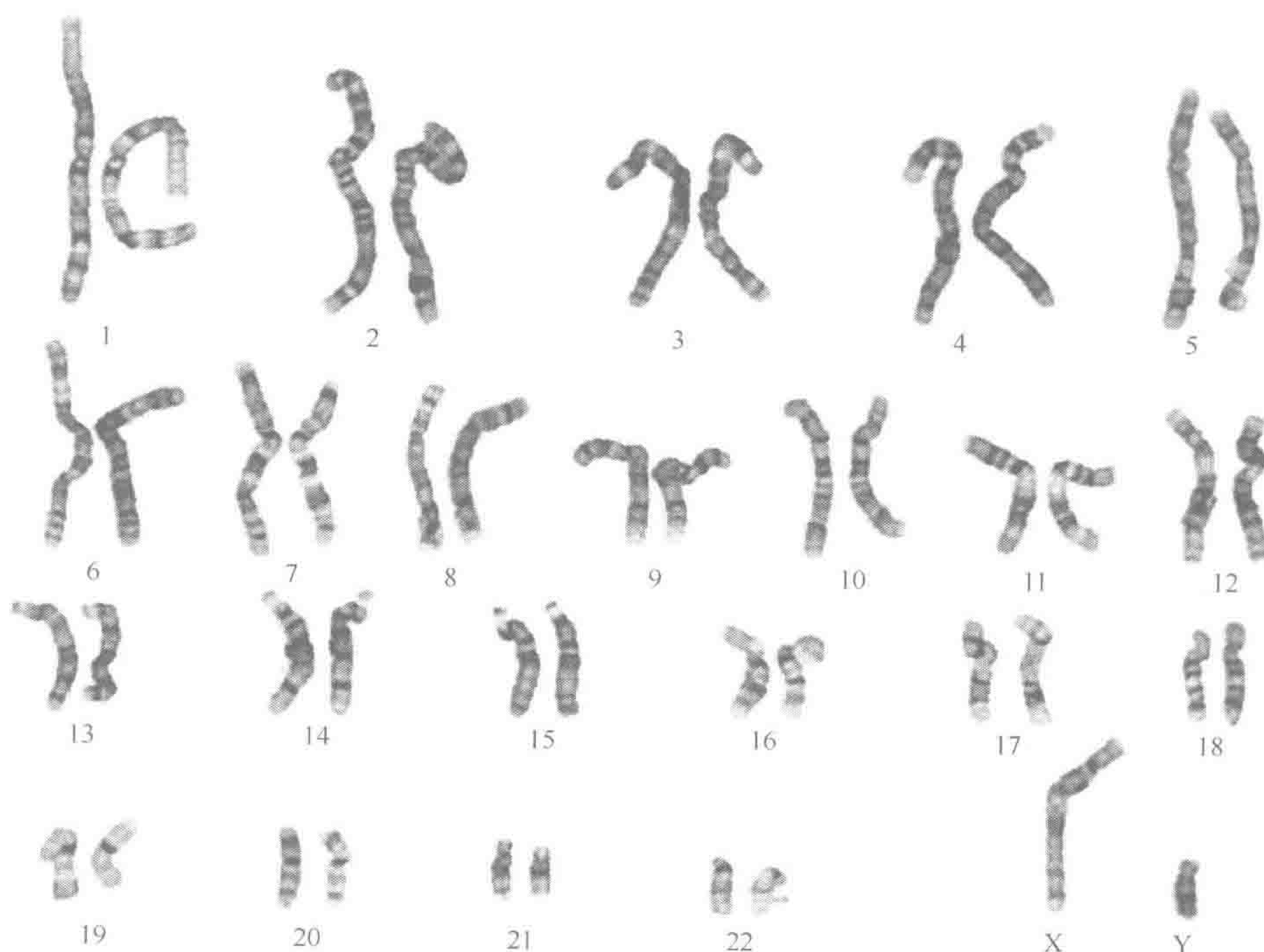


图 2.14　正常男性淋巴细胞有丝分裂染色体的 G 带前中期核型图，单倍体组 550～850 条带

与图 2.15 中理想化的染色体模式图相比，中期染色体的长度介于 2～10μm，如果将细胞中的 DNA 伸直，那么长度约长达 2m。选自 Cross 和 Wolstenholme（2001），人类细胞遗传学：结构分析，3rd Edn（ed. D. E. Rooney），经 Oxford University Press 同意而复制。

类似于 G 带的带可用**喹吖因**（quinacrine）染色产生，该染色剂优先结合到富含 AT 的 DNA 上，而 R-显带可用**色霉素**（chromomycin）染色，因为色霉素优先结合到富含 GC 的 DNA 上。尽管人类 G 带 DNA 的 AT 含量只比 R 带 DNA 的 AT 含量高少数的百分点，G 带却是低 GC 区（即 G 带的%GC 始终比其紧邻的旁侧序列要低）（Niimura and Gojobori，2002）。Saitoh 和 Laemmli（1994）提出，这一差异取决于支架环结构的差异。研究者们认为，在特殊的**支架附着区**（scaffold attachment regions，SAR），染色质环附着到染色体的支架上。依据 Saitoh 和 Laemmli 的模型，G 带中有更小的环且沿着支架 SAR 排列更紧密。这就意味着，相比 R 带，G 带每一 DNA 单位长度中有更多的 SAR，所以用 AT 选择性染色剂会导致更强的着色，就像 Giemsa 那样。

2.4.2　分子细胞遗传学：染色体 FISH

标准的**染色体原位杂交**（chromosome *in situ* hybridization）涉及用标记的 DNA 探针与用空气干燥的显微镜玻片上变性的中期染色体 DNA 标本进行杂交。过去，探针使用放射性同位素来标记，但这种探针会产生较高的噪音信号：信号比率，但开发的**染色体 FISH**（**荧光原位杂交**，fluorescence *in situ* hybridization）（Trask，1991；van Ommen *et al*.,1995）使检测敏感度和分辨率有了显著提高（见图 2.16 基本方法）。在这一技术中，标记 DNA 探针即可直接与标记的荧光素核苷酸前体结合，又可间接的与一个含有报告分子（诸如生物素或地高辛）的核苷酸结合，该分子结合到 DNA 之后，再与荧光标记的亲和力分子结合（节 6.1.2）。为了提高杂交信号的强度，如果可行的话，通常优先使用大的 DNA 探针。

FISH 的优点在于可以快速给出结果。借助**荧光显微镜**（fluorescence microscope）用肉眼即可方便的进行计算（基于荧光显微镜的原理，图 6.5）。便利的染色体 FISH 使用中期的分散的染色体（**中期 FISH**，metaphase FISH），而阳性杂交信号通常显示为双点，正相当于探针与两个姐妹染色单体杂交（图 2.17A）。通过使用精密的图像处理设备以及携带多种不同荧光素的报告结合分子，可以同时对若干 DNA 克隆进行制图与排序。

中期 FISH 最大的分辨率是若干 MB，而更伸展的前中期染色体能有 1Mb 分辨率。但由于染色质折叠的问题，两个差异的标记探针的信号可能连在一起，除非距离大于 2 Mb其信号才能分开。然而，近来人们又研发了新的技术，诸如纤维 FISH 涉及人工伸展 DNA 或染色质纤维，它可达到很高的分辨率（Heiskanen *et al*.,1996）。对间期核染色体进行 FISH（间期 FISH，图 2.17B）也能提供很高的分辨率分析，因为与中期或前中期相比，间期染色体在自然情况下更为伸展。

框 2.3　人类染色体命名法

国际人类细胞遗传学命名系统（ISCN）是由人类细胞遗传学标准委员会确定的。1971 年在巴黎的一次会议上确定了显带染色体的基本术语，这就是经常提到的巴黎命名。

框 2.3 人类染色体命名法（续）

短臂用 p（petit）标记，长臂用 q（queue）标记。每条染色体臂分为区，从着丝粒向外数分别用 p1、p2、p3 等或 q1、q2、q3 等来标记。这些区由特异的界标来限定，这些界标通常是恒定的且具有显著的形态性特征，诸如染色体臂的末端、着丝粒或者一定的带。区分为带标记为 p11（1-1，不是 11）、p12、p13 等，而亚带则标记为 p11.1、p11.2 等，次亚带标记为 p11.21、p11.22，每一种命名都是从着丝粒往外数（图 2.13，图 2.15）。

与着丝粒相对的距离用**近侧**（proximal）和**远侧**（distal）来描述。因此近侧 Xq 即指 X 染色体长臂上最靠近着丝粒的那个染色体片段，而远侧 2p 即指 2 号染色体短臂上离着丝粒最远的那部分，因此也就是离端粒最近的部分。其他常用术语参见下面。

在将人类的染色体和其他物种的染色体比较时，约定使用属名的第一个字母和种的前两个字母来代表（例如 HSA18 即 Human-Homo Sapiens）18 号染色体。

2.4.3 染色体涂染、分子核型分析和比较基因组杂交

标准染色体涂染

FISH 的特殊应用是 DNA 探针的使用。最初的 DNA 是由大量收集的单一类型染色体的不同片段所组成，这样的探针能通过结合插入在一染色体特异 DNA 文库中所有人类 DNA 而制备（节 8.3.2）。另外，可从人类单一染色体杂种细胞中扩增人类特异性片段。一类人-啮齿类的杂种细胞，含有啮齿类全套的染色体和单个类型的人染色体（框 8.4 中杂种细胞制备详解：人的序列能由 Alu-PCR 选择性扩增，一种运用来自 Alu 重复序列引物的 PCR 方法，啮齿类基因组中不含有 Alu 序列，框 5.1）。

产生的杂交信号代表了分布在整条染色体上不同位点的多个标记的 DNA 克隆的结合特性，也就是**染色体涂染**（chromosome paint），使整条染色体都显示荧光［**染色体涂染**：见 Ried 等（1998）和图 2.16 方法基础］。图 2.4 显示间期细胞核中染色体涂染的一个例子，但大多数染色体涂染是在中期染色体上进行。它的一个重要应用就是在临床及癌细胞遗传学中确定新的染色体重排和标记染色体（见图 2.18 的例子）。在肿瘤染色体制备质量差的情况下，它对癌细胞遗传学的研究非常有帮助。

分子核型及多色 FISH（M-FISH）

由于用在区分不同染色体的不同颜色的荧光染料相对较少，所以染色体涂染从一开始便有其局限性。为了增加检测不同靶目标的数量，研究者使用了组合标记（用多种荧光剂标记单个探针）和标记的比率（不同的荧光剂的不同比率）（Lichter，1997）。使用带有专用滤光镜的标准荧光显微镜是无法检测这些混合颜色，相反，研究者更喜欢使用借助多种组合荧光素的人工伪彩色的自动数字图像分析仪。

借助以上的方法，近来可以同时观察到人类 24 条不同的染色体形成的**分子核型**（molecular karyotyping）。一般该方法称作**多色 FISH**（multiplex FISH，M-FISH），并



图 2.15　人类染色体显带模型

这幅图中每条染色体并不是借助显微镜在一个细胞中直接观察到的图像，而是能被观察到的最好的显带模型的汇辑的图像。除了 21 号染色体比 22 号实际上要小之外，其他的染色体均按照大小进行编号。近端着丝粒染色体如 13、14、15、21 和 22 号染色体短臂的重复的核糖体 DNA 基因的排列看似是携带染色质结节的细柄（随体）。组成性异染色质出现在着丝粒、Y 染色体长臂的大部分区域、1q、9q 和 16q 的次缢痕区和近端染色体的短臂。

对使用借助CCD（Charge Coupled Device）相机分别拍摄的五种不同荧光素的数字图像进行分析。图像通过一个软件包进行分析，该软件根据荧光素的组成给予每条染色体一种不同的伪彩色，从而形成一个合成图像。这种方法尤其适用于高频出现的复杂染色体重排的肿瘤样本的分析（见图17.10的例子）。

比较基因组杂交（comparative genome hybridization，CGH）

比较基因组杂交（CGH）是染色体涂染的进一步发展，CGH同时用两种不同的颜色来涂染染色体，获取两个相关来源的总DNA作为探针，此技术可显示亚染色体区甚至整条染色体的扩增或丢失。通常我们将此技术用于肿瘤样本的分析，以找到基因组中已扩增区域或是亚染色体区域已丢失的证据。这两种变化可通过染色体的两种颜色信号的比率的对比来进行判定。

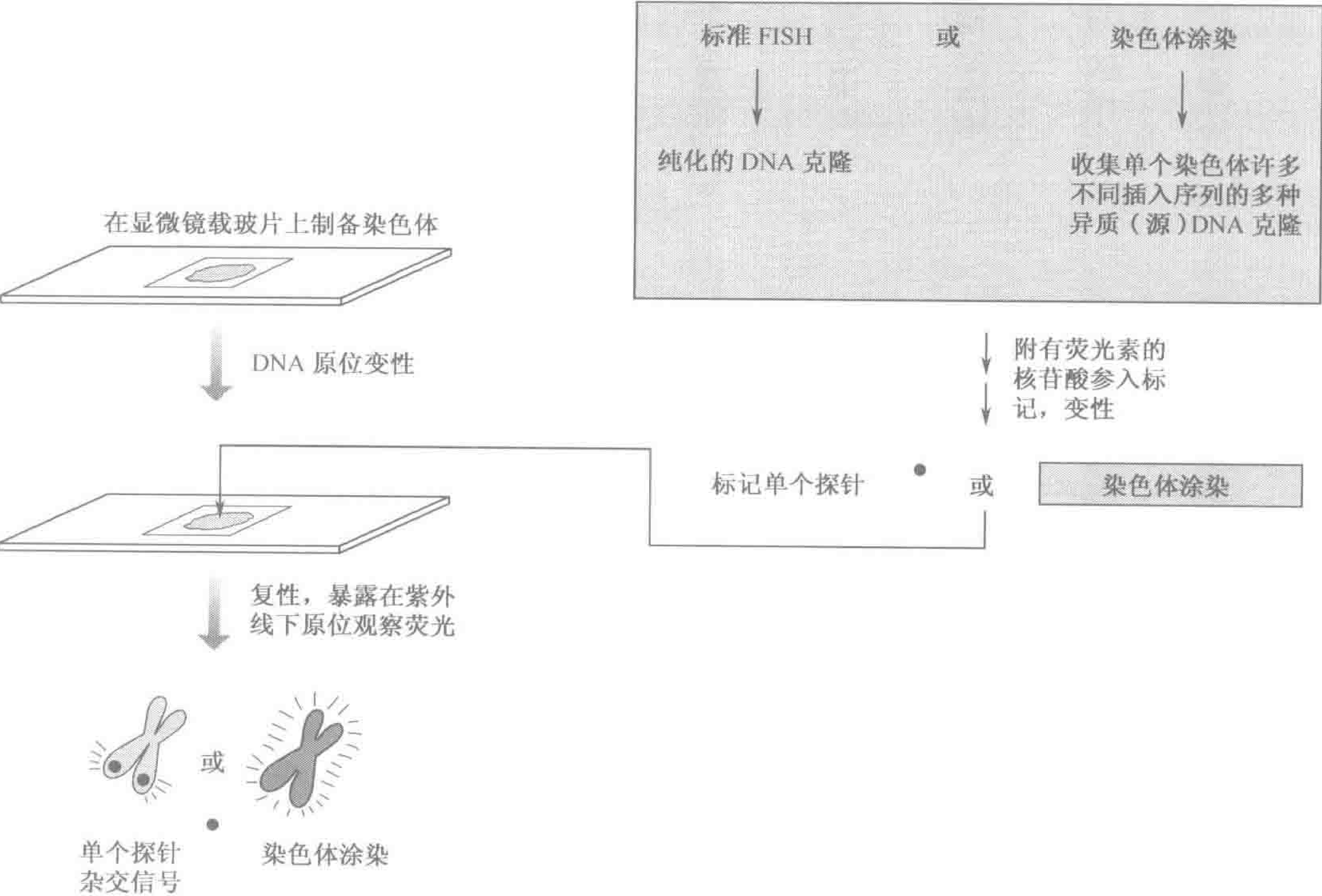


图2.16 染色体FISH和涂染方法的基本原理

注：为了简单起见，只表示了中期染色体的最终结果。中期FISH和间期FISH分别见图2.17A和2.17B，中期染色体涂染及间期染色体涂染分别见图2.18和图2.4。

2.5 染色体畸变

染色体畸变可定义为导致染色体发生可见的改变。能观察到多少取决于技术的应用。用传统的方法来观察标准细胞遗传学样本的最小扩增或缺失大约是4Mb的DNA，然而FISH可观察更小的改变；**分子细胞遗传学**（molecular cytogenetics）的发展已经

扫除了描述为染色体的异常改变和认为是分子或 DNA 的缺陷的改变之间的分界线。染色体畸变定义为由特殊的染色体机制引发的异常。大多数染色体畸变产生的原因主要是由断裂的染色体错误修复，不适当的重组或有丝分裂或减数分裂过程中染色体的不分离引起的。

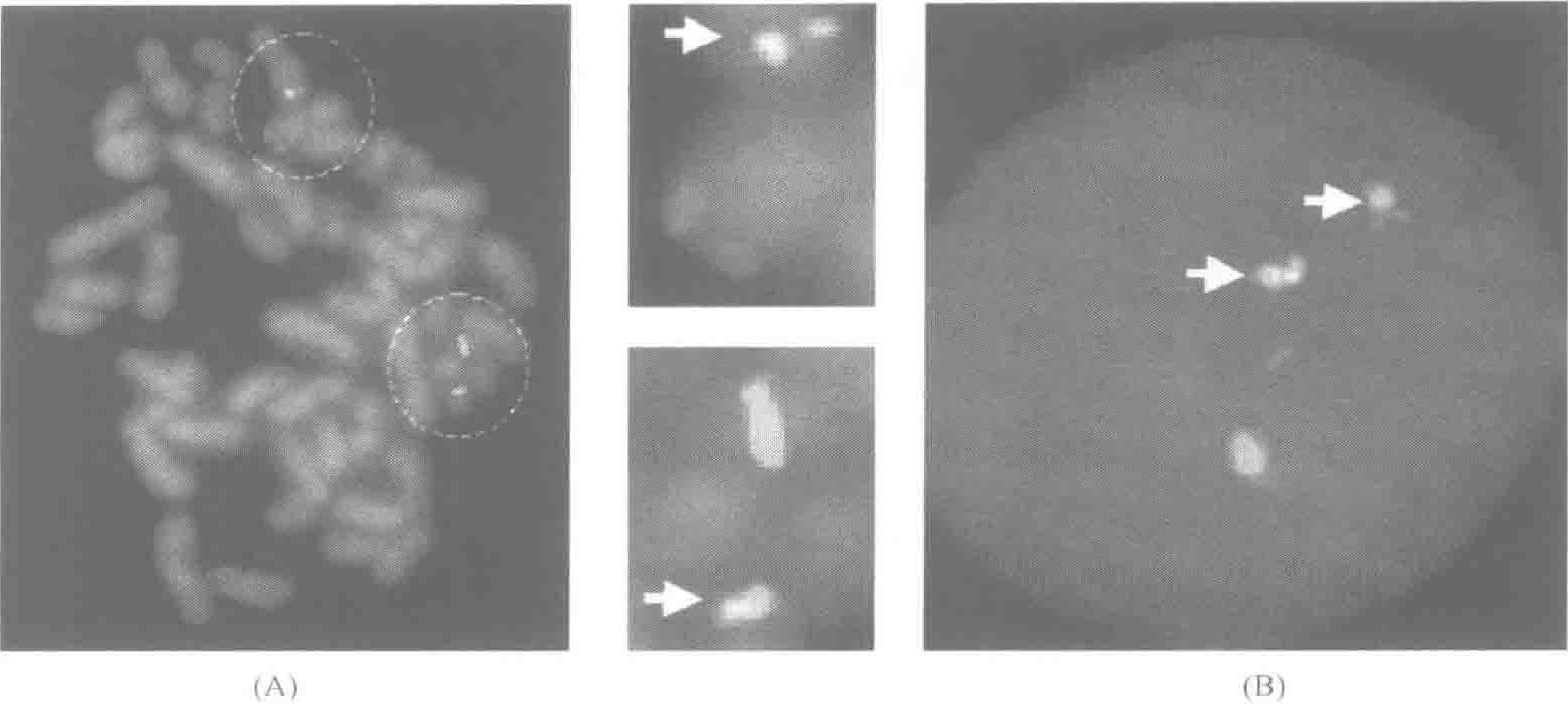


图 2.17　用中期和间期双色 FISH 检测慢性粒细胞性白血病中 BCR-ABL 重排
大多数慢性粒细胞性白血病是由于发生 t（9；22）相互易位，使位于 9q34 的 ABL 癌基因和位于 22q11 的 BCR 基因间的融合（见图 17.4 详解）。此图中表示用标记的 ABL（红信号）和 BCR（绿信号）基因探针杂交显示**慢性粒细胞性白血病患者的中期染色体**（左图圈定的部分，放大为中间图板）和来自**同一患者的间期染色体**的杂交结果（右图板）。正常的 ABL 和 BCR 基因分别显示红色和绿色的信号（**注**：至少可见 ABL 基因在两姐妹染色单体上的中期信号）。白色箭头所指的是在两条易位的染色体［衍生（9）号和（22）号］的 BCR-ABL 融合基因特异信号。由于红色和绿色的信号距离比较近，且红色和绿色信号互相叠加产生了橙黄色信号，故而可以做出判断。图片由 Newcastle Upon Tyne，Institute of Human Genetics 的 Fiona Harding 提供。

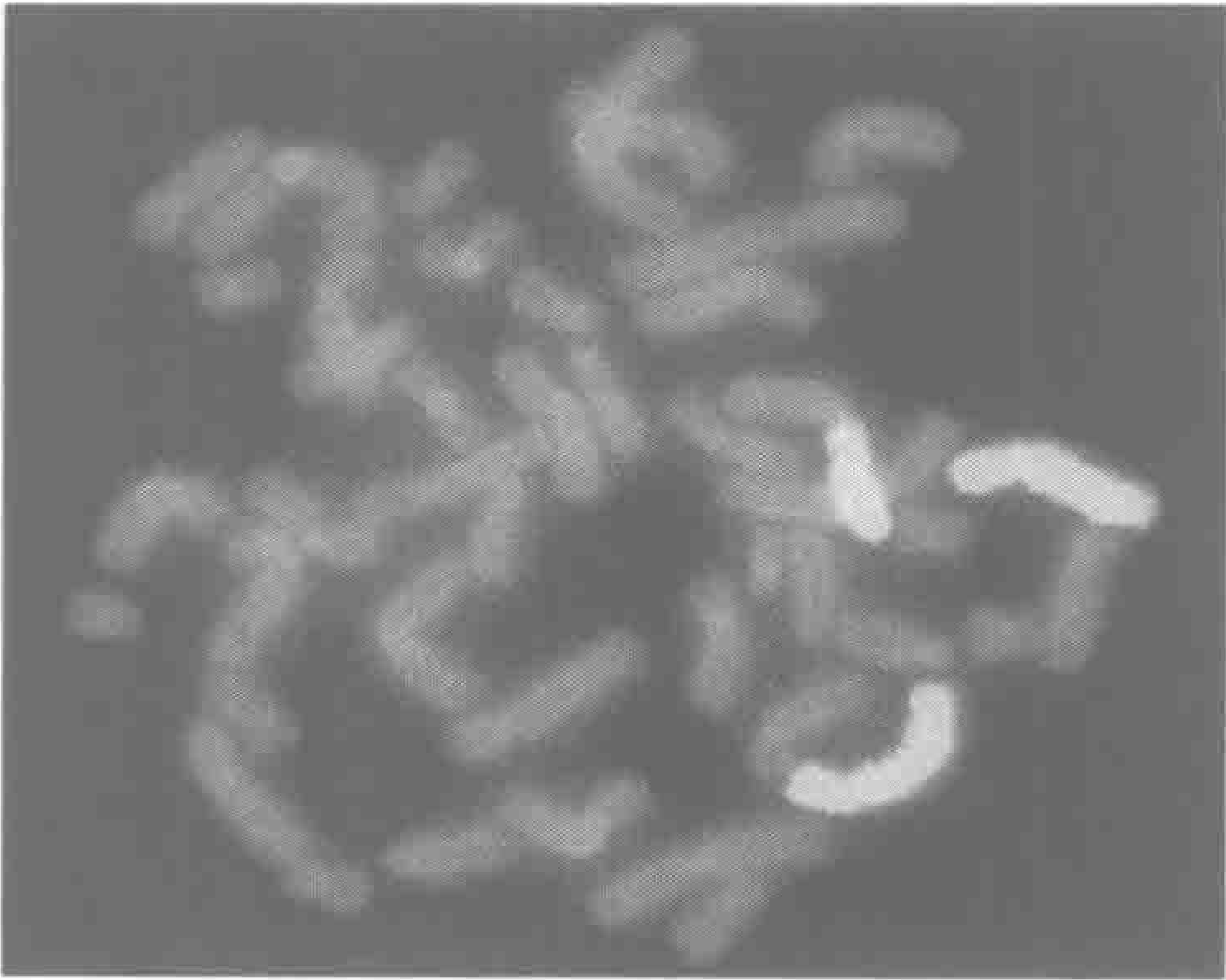


图 2.18　用染色体涂染鉴定染色体重排
一例 X 染色体异常病例，外周血制备的核型证明 X 染色体短臂存在额外的染色体物质，并应用整个 X 染色体涂染（红色）和整个 4 号染色体涂染（绿色）进行了研究。图像证实异常的 X 染色体的短臂上增加的物质来自于 4 号染色体。染色体背景用蓝色的 DAPI 染色。图片由 Newcastle Upon Tyne，Institute of Human Genetics 的 Gareth Breese 提供。

2.5.1　染色体畸变的类型

根据染色体畸变在身体受累细胞的范围可将其分成两类。**组成性异常**（constitutional abnormality）存在于人体所有细胞。该异常一定发生于发育的很早期，很可能是精子、卵子异常，或是异常受精，或是胚胎早期的异常。体细胞（或获得性）异常仅存在于个体某些细胞或组织中。像这样体细胞异常的个体属于**嵌合体**（mosaic）的（图 4.10），含有两种不同染色体组成的细胞，且两种类型的细胞均来自同一合子。

不论是组成性还是体细胞染色体畸变，大多可分成两类：数目和结构异常（框 2.4）。偶尔可见染色体数目和结构均正常的异常，这是由于来自双亲的遗传物质的不平衡（节 2.5.4）。可能是由于双亲遗传物质修正时发生了意外。

2.5.2 染色体数目异常涉及整条染色体的增多或丢失

染色体数目异常可分为三类：多倍体、非整倍体和混倍体。

多倍体

可鉴别的人类妊娠中 1%～3%是**三倍体**（triploid），最常见的原因是两个精子与单个卵子受精（**双受精**，dispermy）；有时是二倍体配子引起的（图 2.19）。三倍体很少能成活，这种情况不适合生存。四倍体更少见且是致死性的。通常是因为没能完成第一次合子分裂：DNA 已经复制其含量为 4*C*，而细胞分裂却未能正常进行。虽然组成性多倍体很少见且是致死性的，但正常人群中也有一些多倍体细胞（节 3.1.4）。

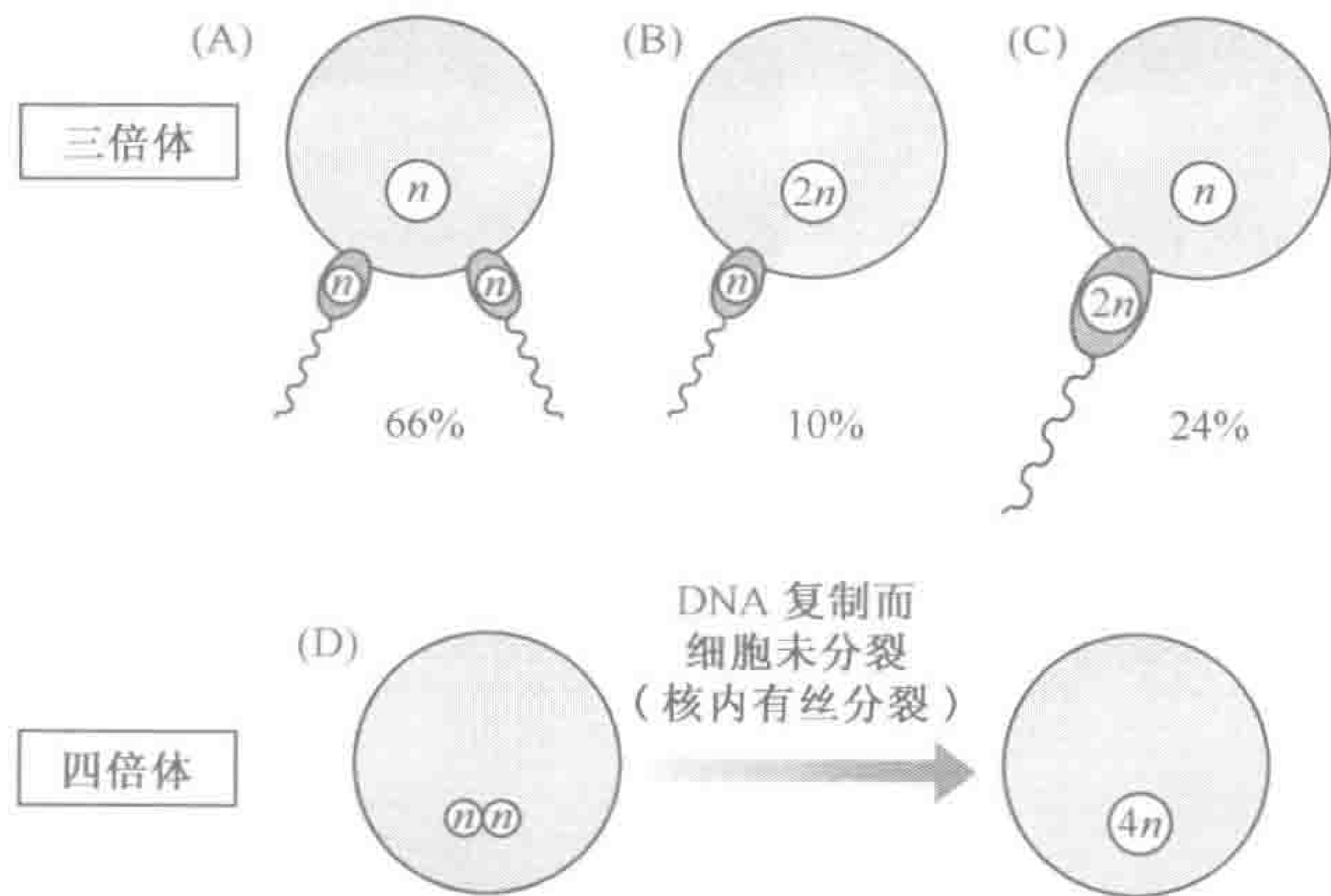


图 2.19 三倍体和四倍体的起源

非整倍体

整倍体（euploidy）是具有完整的染色体组（*n*、2*n*、3*n* 等）。而**非整倍体**（aneuploidy）恰好相反：一个或多个染色体具有一个额外的拷贝或是从整倍体组中丢失一个拷贝。**三体**（trisomy）是指在异常的二倍体细胞中某一特定的染色体有三个拷贝，诸如 Down 综合征中就是出现了 3 条 21 号染色体（47，XX，+21 或 47，XY，+21）。单体是指某一染色体相应丢失，诸如 Turner 综合征中 X 染色体单体（45，X）。癌细胞中常常表现为更多的非整倍体，并伴有多种染色体畸变（图 17.10）。非整倍体细胞的产生主要机制有两种：

- **不分离**（nondisjunction）：配对的染色体在减数分裂后期 Ⅰ 分离失败，或是姐妹染色单体在减数分裂 Ⅱ 或有丝分裂过程中分离失败。减数分裂不分离则会产生含 22 或 24 条染色体的配子，与正常的配子受精后就会产生单体的或是三体的合子。有丝分裂不分离形成嵌合体。

- **后期延缓**（anaphase lag）：由于染色体或染色单体后期的移动迟缓导致不能成功地随着细胞分裂进入到子细胞核中，导致未能进入子细胞核的染色体丢失。

混倍体（嵌合性和嵌合体）

混倍体是指个体中存在着两个或是多个从遗传上不同的细胞系。这些不同的细胞群可产生于同一合子（**镶嵌现象**，mosaicism），或更少见的是起源不同的合子（**嵌合状态**，chimerism），但这样的情况比较少见。详解见节 4.3.6 和图 4.10。组成型的致死性染色体畸变可能会适应在嵌合体中生存。

非整倍体嵌合性（aneuploidy mosaic）常见，例如，有部分正常细胞和部分非整倍体细胞（如三体性）的嵌合性起因于早期胚胎（单体细胞通常死亡）有丝分裂不分离或是后期延迟。

多倍体嵌合性（polyploidy mosaic）（如人二倍体/三倍体嵌合性）偶尔可见。有丝分裂不分离所致的染色体单倍体组增多或丢失不大可能，人类二倍体/三倍体嵌合最有可能是第二极体与一个正常二倍体合子的卵裂核融合而形成的。

染色体数目异常的临床后果

染色体数目异常会带来严重的通常是致死的后果（表 2.4）。尽管 Down 综合征个体中额外的一条 21 号染色体是从正常亲代遗传而来的完全正常的染色体，但它的存在却引起了多种先天性异常。常染色体单体性要比三体性有更加灾难性的后果。这些异常一定是不同染色体上基因编码产物水平不平衡的结果。正常生长发育和功能的实现取决于基因产物间的无数的相互作用，包括许多不同染色体上基因编码产物。改变相关染色体数目将影响到这些相互作用。

表 2.4　染色体数目异常的后果

多倍体	
三倍体（69，XXX，XXY 或 XYY）	发生率 1%～3%；几乎没有能活着出生的，不能成活
非整倍体（常染色体）	
缺倍体（丢失一对同源染色体）	植入前致死
单体性（丢失一条染色体）	胚胎致死
三体性（一条额外染色体）	通常在胚胎期或胎儿期死亡，但是 13 三体（Patau 综合征）和 18 三体（Edwards 综合征）可能存活至足月，21 三体（Down 综合征）可能活到 40 岁或更长。
非整倍体（性染色体）	
额外的染色体	（47，XXX，47，XXY 或 47，XYY）表现的症状相对较小，寿命正常
缺少的染色体	45，X=Turner 综合征，约 99%的病例自然性流产，成活下来的智力正常但不能生育且有轻微体征改变。45，Y=不能成活。

性染色体数目异常致病性比常染色体数目异常要少得多。具有 47，XXX 和 47，XXY 核型的人，其机体的功能常常在正常范围内，与具有任何常染色体三体和单体的

个体相比，47，XXY 的男性受累程度相对较小；45，X 的女性患者主要症状明显要轻。事实上，正常人可以有一条或两条 X 染色体，有一条 Y 或没有 Y 染色体，性染色体的可变数目一定有特殊的机制使之发挥正常的功能。以 Y 染色体为例，因为它携带非常少的基因，它的重要功能是决定男性的性别。对于 X 染色体，在哺乳动物中存在着 **X 染色体失活**（X-chromosome inactivation）的特殊机制（节 10.5.6），X 染色体编码的基因产物的水平调节不取决于细胞中 X 染色体的数目。

框 2.4 染色体畸变命名法

数目异常：

三倍体 69，XXX、69，XXY、69，XYY

三体性 例如 47，XX，+21[a]

单体性 例如 45，X

嵌合型 例如 47，XXX/46，XX

结构异常：

缺失 例如 46，XY，del（4）（p16. 3）[b]；46，XX，del（5）（q13q33）[b]

倒位 例如 46，XY，inv（11）（p11p15）

重复 例如 46，XX，dup（1）（q22q25）

插入 例如 46，XX，ins（2）（p13q21q31）[c]

环状染色体 例如 46，XY，r（7）（p22q36）

标记染色体 例如 47，XX，+mar[d]

易位，相互易位 例如 46，XX，t（2；6）（q35；p21. 3）[e]

易位，罗伯逊易位 例如 45，XY，der（14；21）（q10，q10）[f]

（产生衍生染色体 ）46，XX，der（14；21）（q10，q10），+21[g]

注：

a. 染色体增多用“+”表示；染色体丢失用“−”表示

b. 末端缺失（断裂点在 4p16. 3），中间缺失（5q13−q33）

c. 2q13−q31 片段插入断裂点 2p13 导致 2 号染色体一个拷贝重排

d. 细胞含有一个**标记染色体**（marker chromosome）的核型（指额外未确定的染色体）

e. 在 2q35 和 6p21. 3 两断裂点的平衡相互易位

f. 14；21 平衡易位的携带者罗伯逊易位。q10 不是真正的染色体带，而是指着丝粒；der 指**衍生染色体**（derivative chromosome）（在一个染色体发生易位时使用）

g. 易位型 Down 综合征；患者有一个正常的 14 号染色体，一个 14；21 罗伯逊易位的染色体和两个正常的 21 号染色体拷贝

这是一个简短的命名；ISCN 表明了更加复杂的命名法则以对任何一种染色体的异常进行完整的描述。见进一步阅读。

在生命的发育最早期阶段，常染色体单体性一定是致死的，每一条染色体上可有少数基因的产物水平降低 50% 时与发育是不相容的。对大多数基因来说，这种减少并没有显著的致病性，它们是微效的，但是成百上千的微效累加足以破坏胚胎的正常发育。在基因产物水平上三体性改变要比单体性改变小，同时它们的效应有点少，故三体性胚

胎存活时间较单体胚胎长。13、18 或 21 三体的胚胎适合存活直至出生。有趣的是，这三种染色体相对来说含有比较少的基因（节 9.1.4）。

三倍体在人和其他动物上是致死的原因并不明确。每个常染色体均有三个拷贝，常染色体的基因剂量也平衡，对此不应该产生问题，但三倍体在减数分裂中 3 条染色体不能配对与正确的分离，导致三倍体总是不育的。而在其他方面，许多三倍体的植物却是健康且有活力，而三倍体动物是致死的，最可能的解释是 X 染色体和常染色体基因编码产物间的不平衡，因为 X 染色体的失活是不能代偿的。

2.5.3　染色体断裂的错误修复或重组系统障碍引起染色体结构异常

染色体断裂的原因是 DNA 的损伤（诸如辐射或化学物）或是由重组的部分机制引起的。在细胞周期的 G_2 期（图 2.1），染色体由两个染色单体组成。在这一阶段发生的断裂为**染色单体断裂**（chromatid break），只影响两条姐妹染色单体中的一条。G_1 期发生的断裂，如果不能在 S 期之前得以修复，将形成染色体断裂而影响到两条染色单体。细胞具有识别的酶系统，如果可能修复的话，断裂的染色体就修复了。修复的方式可以是将两个断裂末端连接在一起或为一个断裂末端加上端粒。正常情况下，细胞周期检控点机制（节 18.7.3）能阻止未修复的染色体进入有丝分裂期；若损伤不能修复，细胞则自发性死亡（凋亡）。

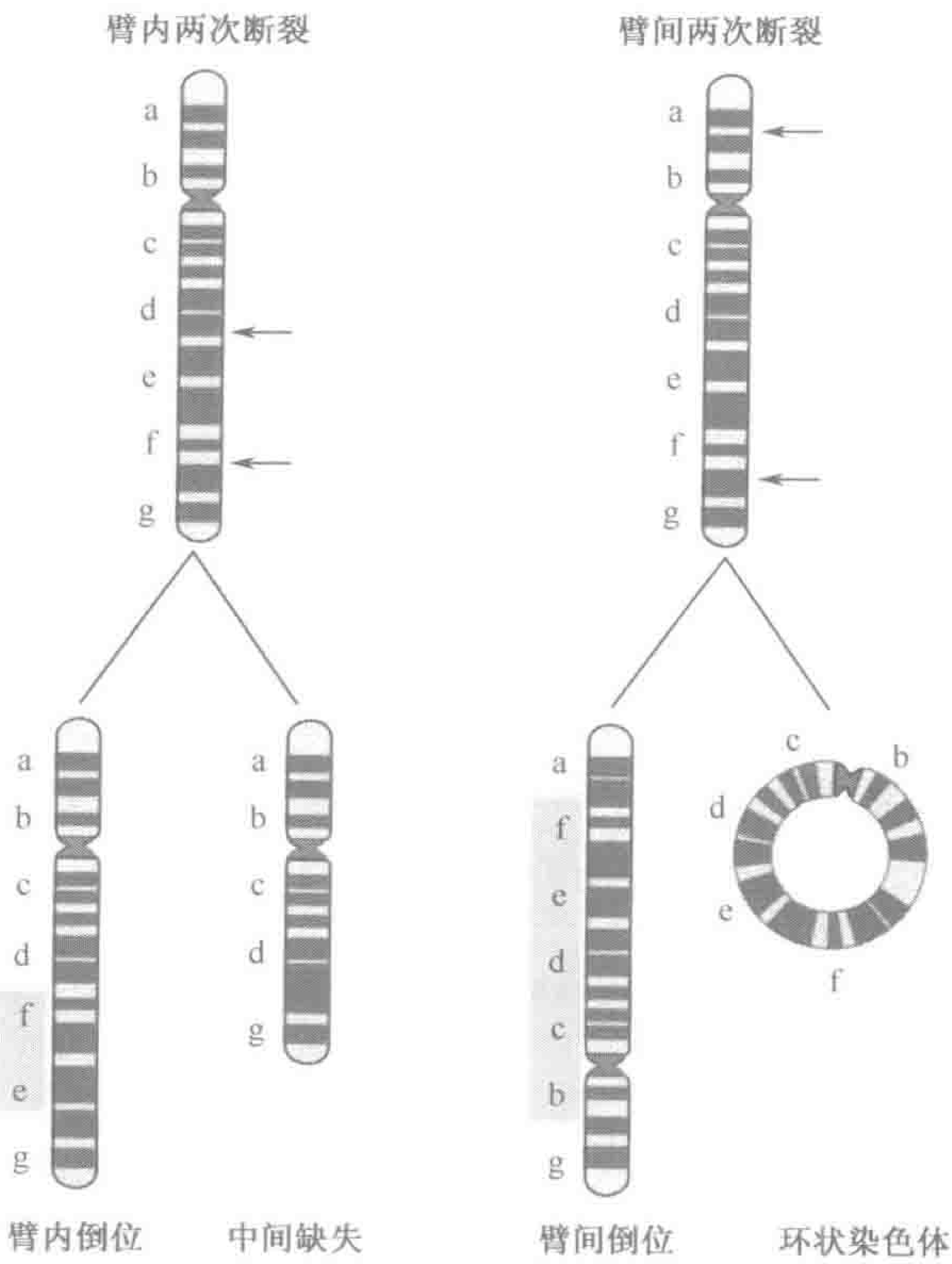


图 2.20　在单一染色体上有两个断裂点可能稳定的结果

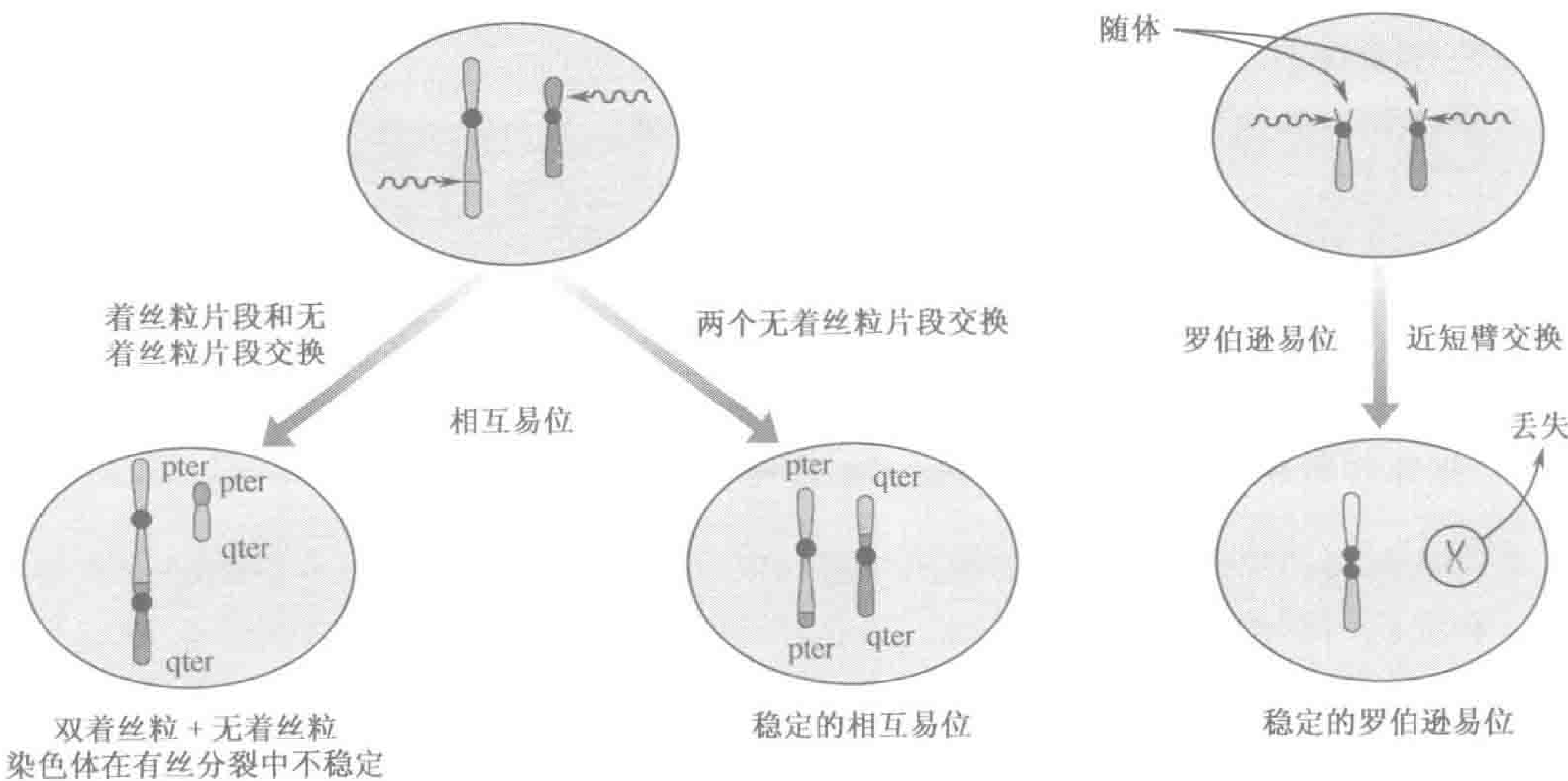


图 2.21　易位的起因

双着丝粒和无着丝粒的染色体在有丝分裂过程中不稳定。罗伯逊易位是由近端着丝粒染色体 13、14、15、21 和 22 号染色体的短臂近端之间的交换产生的［在近端着丝粒染色体 13、14、15、21 和 22 号的短臂排列重复 rDNA基因，常常显示细丝状连有染色质球状小体-**随体**（satellite）］。罗伯逊易位中出现两个着丝粒，但它们的功能作为一个着丝粒而染色体是稳定的。小的无着丝粒片段丢失，但这并不会产生致病性后果，因为此片段中仅包含重复 rDNA 序列，而此序列同样也存在于其他近端着丝粒染色体上。

当染色体的断裂不能正确修复时则发生结构异常。假如没有游离的断裂末端，它就有可能通过细胞周期检控点。因此，有时断裂末端错误的连接在一起。任何**无着丝粒的染色体**（acentric chromosome）（缺少着丝粒）或**双着丝粒的染色体**（dicentric chromosome）（含两个着丝粒）在有丝分裂过程中不能稳定的分离，终将丢失。然而那些单个着丝粒的染色体即使它们结构异常也可以在连续的有丝分裂循环中增殖。减数分裂过程中错配染色体间的重组是易位常见的原因，尤其是在精子发生中。表 2.5 归纳了主要的稳定的结构异常。

表 2.5 中未列出另一类少见的染色体结构异常是**等臂染色体**（isochromosome）：某一特定的染色体两个长臂或者是两个短臂形成的对称的染色体。它们被认为是由于一条染色体邻近着丝粒处姐妹染色单体间发生了异常的 U 型交换。除 i（Xq）和 Down 综合征偶尔发生 i（21q）外，人类等臂染色体罕见。

表 2.5　染色体断裂错误修复或非同源染色体间重组引起的染色体结构异常

	涉及一条染色体	涉及两条染色体
一次断裂	末端缺失（通过添加端粒来修复）	
二次断裂	中间缺失；倒位	相互性易位（图 2.21）
	环状染色体（图 2.20）	罗伯逊易位（图 2.21）
	不对称姐妹单体交换重复或缺失（图 11.7）	不相等重组引起重复或缺失（图 11.7）
三次断裂	多种重排如倒位伴缺失，染色体内插入	染色体内插入（直接或间接）

如果没有真正染色体物质的获得或丢失，则染色体结构异常为**平衡性**（balanced）的，若有获得或丢失则为**不平衡**（unbalanced）。平衡性的异常（倒位，平衡易位）对表现型没有影响，以下情况除外：

▶ 染色体断裂可能破坏一个重要基因；

▶ 即使不破坏基因的编码序列，断裂可能会影响基因的表达，这可能使一个基因脱离了某个调控元件的控制，或是使该基因处于一个不适宜其表达的染色质环境，例如一个正常有活性的基因易位至异染色质；

▶ X 染色体与常染色体间的平衡易位会引起 X 染色体的失活问题（图 14.10）。

有时会把罗伯逊易位称为**着丝粒融合**（centric fusion）。但这是误解，事实上断裂是发生在短臂近端。易位染色体确实是双着丝粒，由于两个着丝粒相距很近，二者作为一个着丝粒发挥作用，分裂时染色体正常分离，短臂末端部分由于是一无着丝粒片段而在分裂过程中丢失。近端着丝粒染色体的短臂只含有重复的 rRNA 基因序列，故两短臂的丢失也无表现型效应。正是因为没有表现型效应，即使在罗伯逊易位中有染色体物质的丢失也被认为是平衡的。

不平衡异常的发生可以是由缺失或少见的重复直接引起，或是由平衡结构异常携带

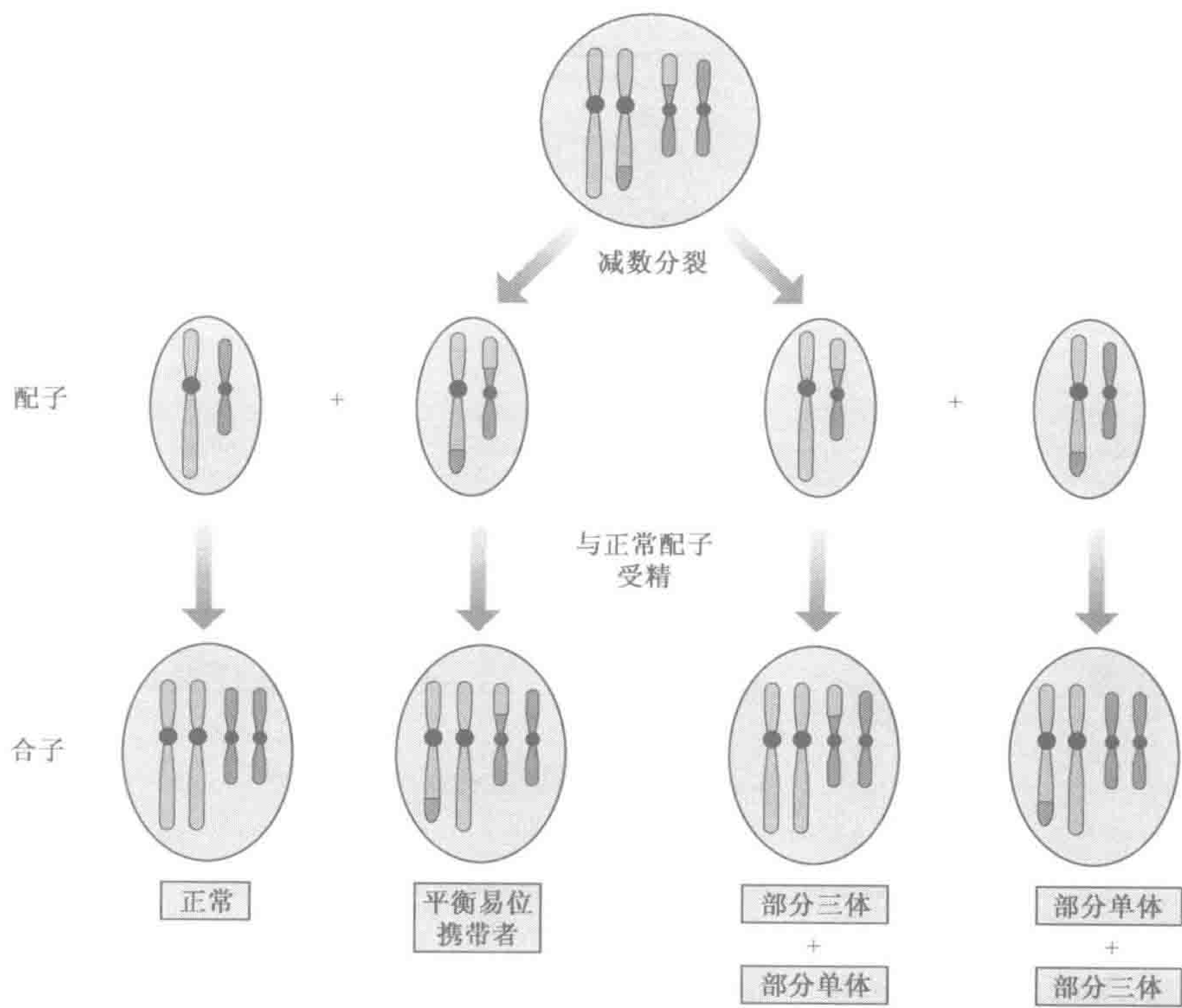


图 2.22　在平衡相互易位携带者的减数分裂结果

其他的分离模式也有可能发生，比如说 3∶1 分离。每种可能的配子出现的相对几率还不能够明确的预测出来。携带者的后代出现某种可能性后果的几率取决携带者配子出现的频率以及异常孕体发育足月的拟然性。更多讨论参考 Gardner 和 Sutherland（进一步阅读）的书。

者在减数分裂过程发生染色体错误分离而间接引起的。如果染色体同源配对结构不相匹配，平衡结构异常的携带者在减数分裂过程中出现下列问题：

- 一个平衡相互易位携带者所产生的配子在受精后所生出的完全正常的孩子，一个表型正常的平衡易位携带者，或多种不平衡核型，即通常有一个染色体的部分单体和另一染色体的部分三体的结合的后代（图 2.22）。人们尚不能对每种结果的相关频率作出一般的表述。任何不平衡片段的大小取决于断裂点的位置，如果不平衡片段大，胎儿可能会自然流产，片段小会生出活的畸形婴儿。
- 一个平衡罗伯逊易位的携带者产生的配子，受精后所生出的后代可以是完全正常的，可以是表型正常的平衡易位携带者，或是一条染色体完全单体或完全三体受累的后代（图 2.23）。

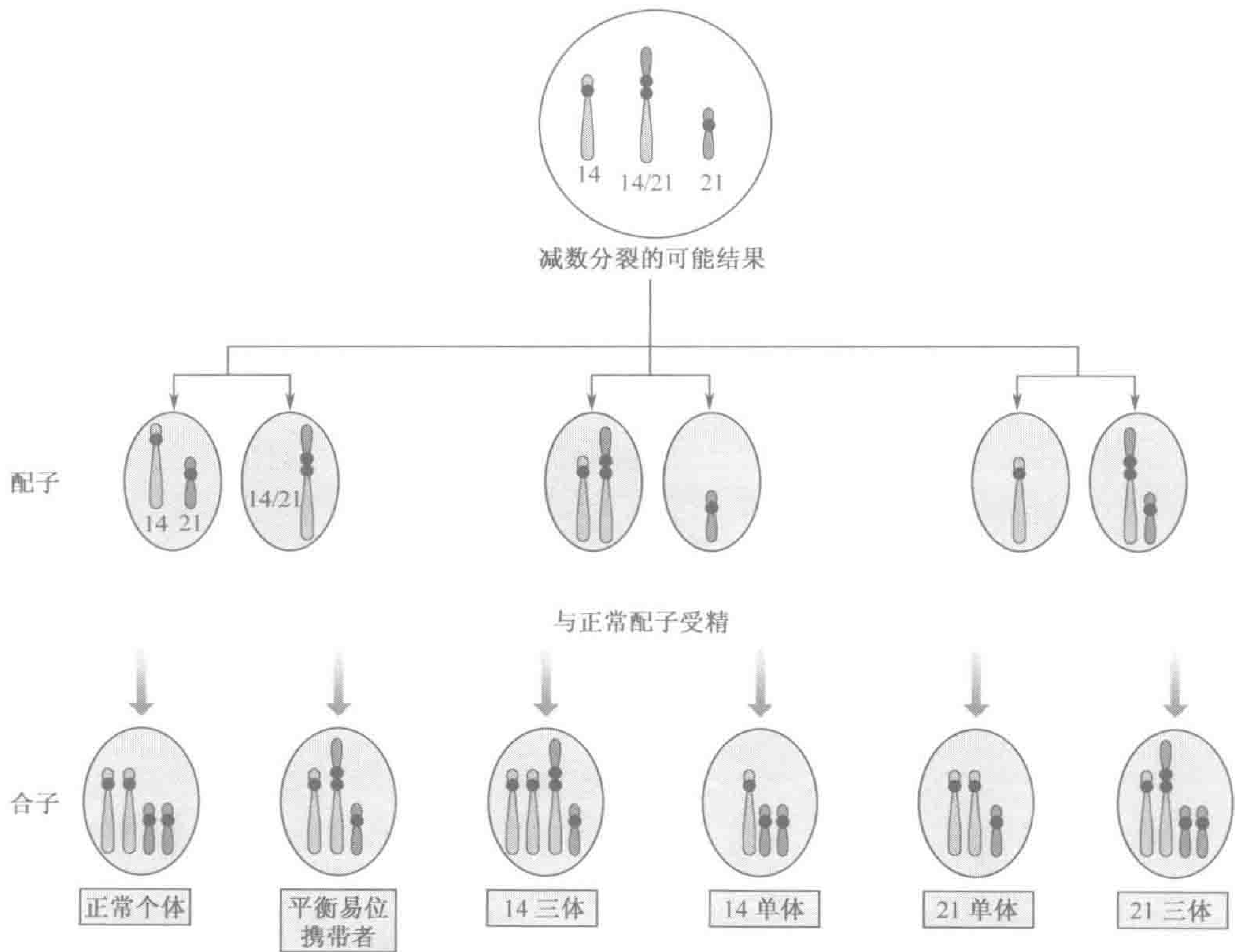


图 2.23　一个罗伯逊易位携带者减数分裂的结果

携带者通常是没有症状，但是携带者会产生不平衡的配子，这些配子会形成单体型或三体型合子。此图例中框内的单体的或三体的合子均不能发育到足月。

- 一个臂间倒位携带者可能会生出不平衡的后代，因为倒位和未倒位的染色体同源配对时会形成一个环以致沿此染色体全长进行片段配对。如果在环内发生交换就会产生一个携带不平衡缺失和重复的染色体。臂内的倒位会形成相似的环形结构，但环内发生的任何交换都会产生一个无着丝粒或双着丝粒的染色体，它是不能成活的。进一步了解倒位携带者的减数分裂细节，参考 Gardner 和 Sutherland（进一步阅读）

的书或其他细胞遗传学教科书。

2.5.4　如果它们具有错误的双亲来源，表现正常的染色体组成可能是致病的

下面讲述的罕见的异常证明，仅有正确的染色体数目和结构是不够的，还必须有正确的双亲来源。两个基因组均来自同一亲代的核型 46，XX 孕体（**单亲二倍体**，uniparental diploidy）决不能正常发育，某些个体两条同源染色体均来自同一亲代（**单亲二体性**，uniparental disomy）同样会引起异常。少数基因在亲代已被印记（节 10.5.4），这些基因按其来源而表达不同。推测单亲二体和单亲二倍体的异常就是由这些印记基因的异常表达引起的。

在**葡萄胎**（hydatidiform mole）中可见单亲二倍体，是一种来自同一亲代核型 46，XX 的异常孕体。葡萄胎妊娠表现为广泛的滋养层增生且没有胎儿部分，葡萄胎妊娠具有转化绒毛膜癌的高风险。遗传标记研究表明，大多数胎块在所有基因座都是纯合性的，这表明，这些胎块都是由单一的精子细胞的染色体加倍而产生的。**卵巢畸胎瘤**（ovarian teratoma）是由母体单亲二倍体形成的。这些少见的卵巢良性肿瘤是由紊乱的胚胎组织构成，没有额外的胚膜。这些肿瘤是由未排出的卵母细胞激活所形成的。

单亲二体性（UPD）受累的一单对同源染色体，如果没有异常发生将是无法诊断，但是当发生特异性综合征的时候（框 16.6），可检测染色体进行鉴别。UPD 涉及**同源二体性**（isodisomy），指两条同源染色体完全相同；或**异源二体性**（heterodisomy），两条同源染色体来自双亲之一。通常认为三体性复原是导致 UPD 的原因：一个孕体要么是三体性，否则就会死亡，偶尔情况下，通过有丝分裂不分离或后期延迟丢失一条染色体而形成多能细胞。整倍体子代细胞形成胚胎而其他所有的非整倍体细胞均死亡。假设这三个拷贝中的任一条被丢失的机会均等，那么单一染色体就会有三分之二的机会丢失而导致正常染色体的组成，而有三分之一的机会是形成单亲二体性（要么来自父体，要么来自母体）。单亲同源二体性可能由于对单体性胚胎的选择压力通过对单体性染色体的选择性复制而形成整倍体而发生的。

（李福才　译）

进一步阅读

Choo KH (2001) Domain organization at the centromere and neocentromere. *Developmental Cell* **1**, 165–177.

Gardner RJM, Sutherland GR (1996) *Chromosome Abnormalities and Genetic Counseling*, 2nd Edn. OUP, Oxford [A thorough introduction to the nature, origin and consequences of human chromosomal abnormalities].

ISCN (1995) *An International System for Human Cytogenetic Nomenclature* (ed. F Mittelman). Karger, Basel.

Pollard TD, Earnshaw WC (2002) *Cell Biology*. Saunders, Philadelphia.

Rooney DE (2001) (ed.) *Human Cytogenetics: Constitutional Analysis*, 3rd Edn. Oxford University Press, Oxford [Detailed laboratory protocols].

Rooney DE (2001) (ed.) *Human Cytogenetics: Malignancy and Acquired Abnormalities*, 3rd Edn. Oxford University Press, Oxford [Detailed laboratory protocols].

Schinzel A (2001) *Catalogue of Unbalanced Chromosome Aberrations in Man*. Walter de Gruyter, Berlin.

Therman E, Susman M (1992) *Human Chromosomes: Structure, Behavior and Effects*, 3rd Edn. Springer, New York [An excellent compact introduction; emphasis on scientific bases rather than clinical implications].

Web-based cytogenetic resources. See compilations such as at: www.kumc.edu/gec/prof/cytogene.html

参考文献

Blackburn EH (2001) Switching and signaling at the telomere. *Cell* **106**, 661–673.

Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.* **10**, 211–219.

Craig JM, Earnshaw WC, Vagnarelli P (1999) Mammalian centromeres: DNA sequence, protein composition, and role in cell cycle progression. *Exp. Cell Res.* **246**, 249–262.

Craig JM, Bickmore WA (1993) Chromosome bands – flavours to savour. *Bioessays* **15**, 349–354.

Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Rev. Genet.* **2**, 292–301.

Cross I, Wolstenholme J (2001) An introduction to human chromosomes and their analysis. In: *Human Cytogenetics: Constitutional Analysis*, 3rd Edn (ed. DE Rooney) Oxford University Press, Oxford.

Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Kohler N (1999) Discovery of tetraploidy in a mammal. *Nature* **401**, 341.

Gilbert DM (2001) Making sense of eukaryotic replication origins. *Science* **294**, 96–100.

Heiskanen M, Peltonen L, Palotie A (1996) Visual mapping by high resolution FISH. *Trends Genet.* **12**, 379–384.

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102.

Huxley C (1997) Mammalian artificial chromosomes and chromosome transgenics. *Trends Genet.* **13**, 345–347.

Manuelidis L (1990) A view of interphase chromosomes. *Science* **250**, 1533–1540.

Niimura Y, Gojobori T (2002) *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc. Natl Acad. Sci. USA* **99**, 797–802.

Parada LA, Misteli T (2002) Chromosome positioning in the interphase nucleus. *Trends Cell Biol.* **12**, 425–432.

Saitoh Y, Laemmli UK (1994) Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold. *Cell* **76**, 609–622.

Schindelhauer D (1999) Construction of mammalian artificial chromosomes: prospects for defining an optimal centromere. *Bioessays* **21**, 76–83.

Trask BJ (2002) Human cytogenetics: 46 chromosomes, 46 years and counting. *Nature Rev. Genet.* **3**, 769–778.

第 3 章　细胞和发育

本章内容

除病毒外，所有的生命都由**细胞**（cell）——含水的、彼此之间及与环境之间相互作用的、膜包裹的区室——所组成的。每个细胞都由存在的细胞分裂或融合产生，最终必然有一条完整的细胞链可以追溯至可能生活于 35 亿年前的最早的成功原始细胞。细胞如何形成是一个有趣的问题。

有些细胞是独立的单细胞生物。这种生物必须完成维持生命所需的所有活动，也必须具有繁殖能力。因此，它们对环境的改变极为敏感，典型地具有非常短的生命周期以适应快速的增殖。结果是，它们能快速适应周围的改变——更能在特殊环境中存活的突变体可迅速兴旺。这导致了单细胞生物的巨大范围，它们已通过进化适应了不同的、有时是极端环境的微生态。然而，这些有机体的复杂性总是有限的。

多细胞生物具有相对较长的寿命，并且表型的改变也相应较慢。它们的成功是基于

将不同的功能划分给不同的细胞种类，大量的细胞-细胞相互作用类型产生的可用性为功能复杂性提供巨大的潜能。有些单细胞生物如黏土霉菌（*Dictyostelium discoideum*），在其生命周期中有一多细胞阶段，但主要以单细胞存在。与之不同的，动物和植物在它们整个生长发育生存期都是多细胞，但留有特殊的**生殖细胞**（germ cell）以利于繁殖。多细胞生物在大小、类型和细胞数目上都可以有巨大的改变，但任何情况下生命都由单细胞开始。从单细胞到成熟有机体，**发育**（development）的过程包括许多次细胞分裂，在此期间细胞必须以精确的模式逐渐地特化和组织。它们在发育期间的行为和相互作用塑造了该有机体的整个形态。

3.1　细胞的结构和多样性

3.1.1　原核生物和真核生物代表细胞生命形式的基本类别

根据细胞早期进化的趋异引起的内部结构和主要功能的差异，细胞可以分为广泛的分类学组（图 3.1）。生物体主要分为原核生物和真核生物，是基于细胞结构的根本差异（图 3.2）。

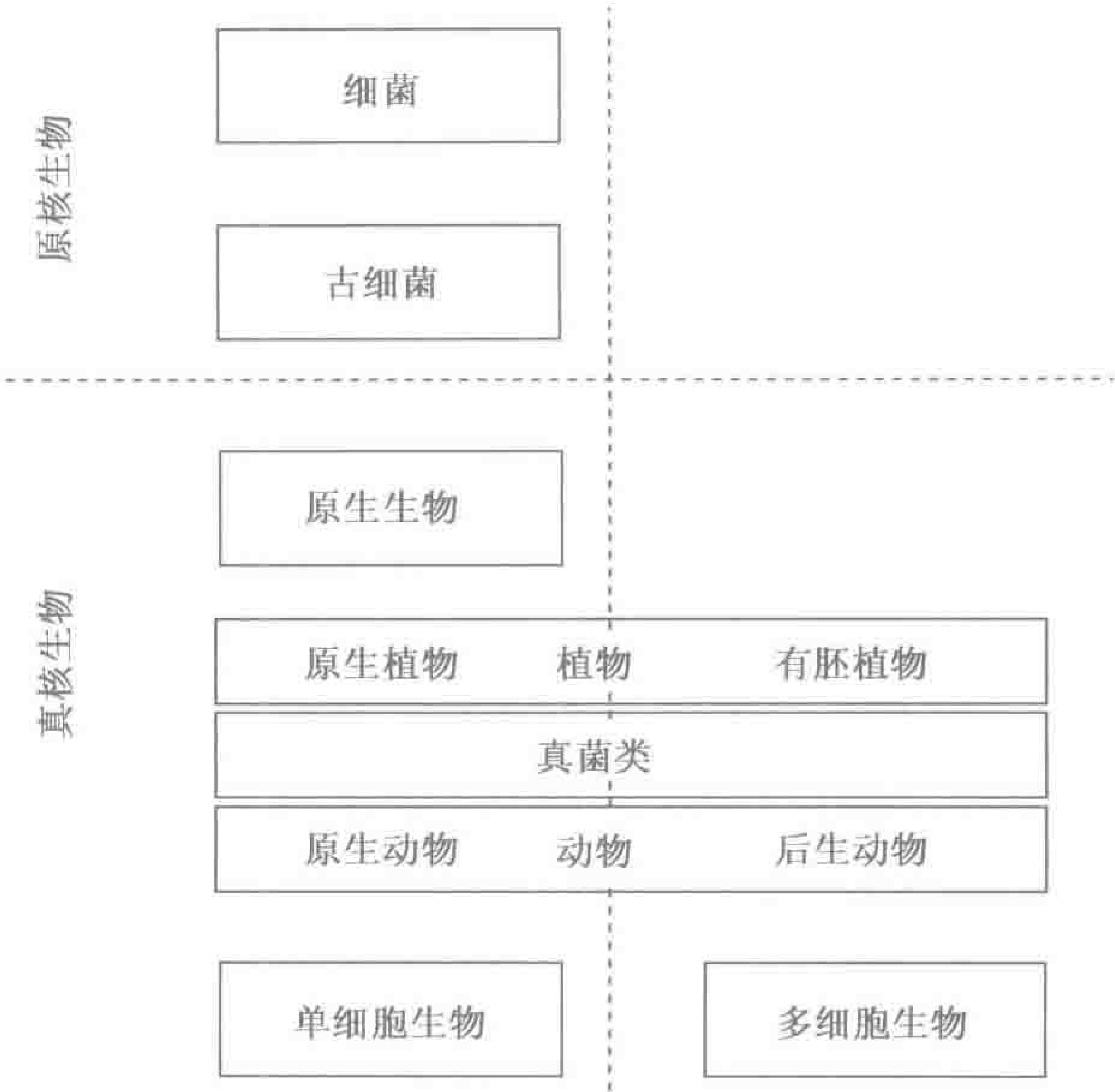


图 3.1　单细胞和多细胞生物的分类

注：原生生物包括分类为动物（原虫）、植物（原生植物）和真菌（如酵母）的单细胞生物，但是许多原生生物并不属于它们任何一组。

原核生物细胞（prokaryote cell）具有简单的内部结构，特别是缺乏细胞器和细胞内间格。它们没有明确的核。染色体 DNA 没有高度的组织，而以核酸-蛋白复合体的形式存在，称为**类核**（nucleoid）。在电子显微镜下，原核生物的细胞显示相对的无特征。然而，因为原核生物经历比我们自身多得多的世代进化，所以它们绝不是原始的。它们由两个生物界组成：**细菌**（nucleoid）［以前称作**真细菌**（eubacteria）以区分于古

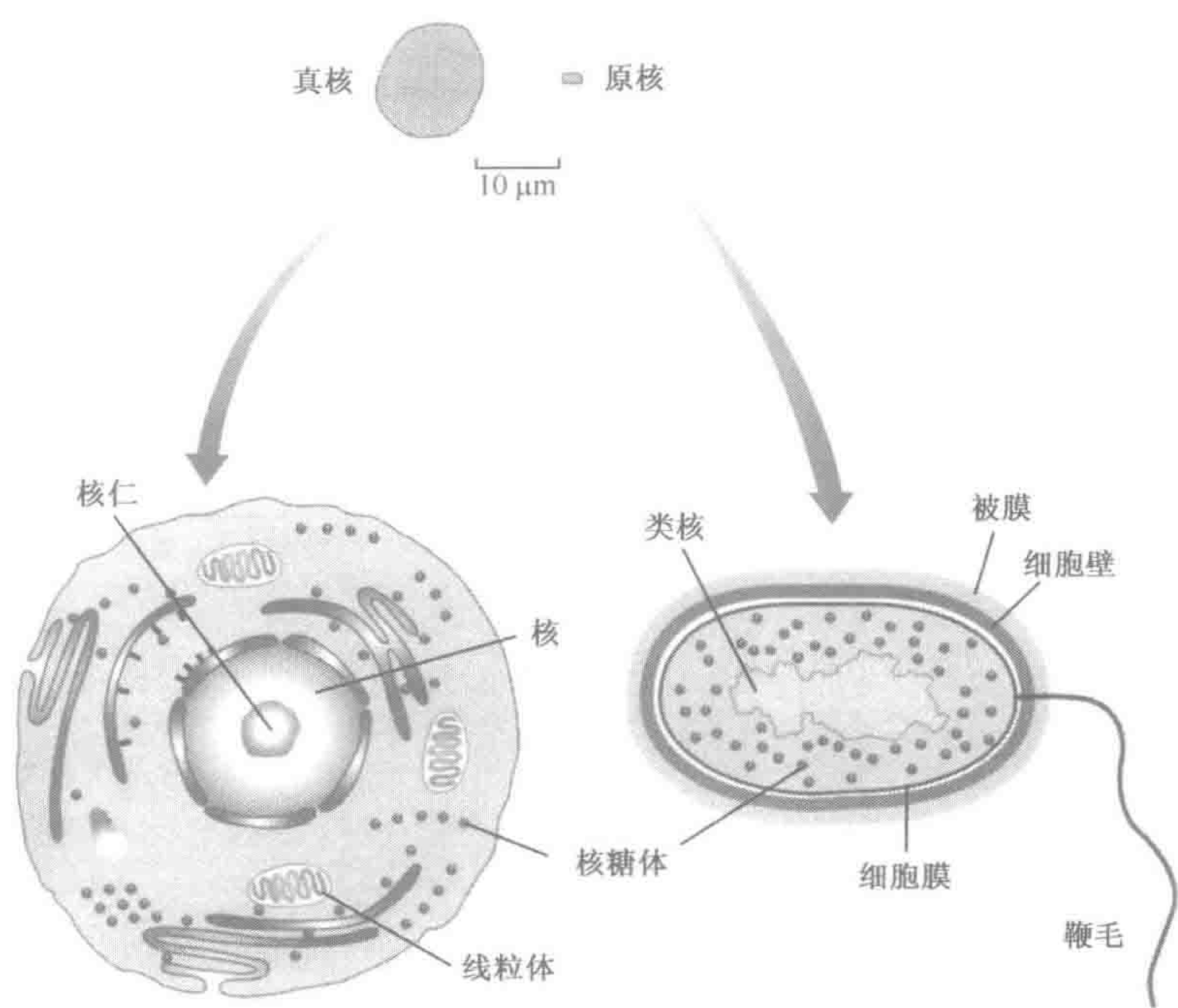


图 3.2　原核和真核细胞解剖学

图中所示真核细胞是一个普通的动物细胞。个别细胞类型的例子在框 3.6 中介绍。

细菌] 和**古细菌** (archaea) [一个了解相当少的生物群体，与细菌表面相似，故以前命名为**原始细菌** (archaebacteria)]。所有原核生物都是单细胞的。细菌可见于许多环境，其中一些对人类是致病的。古细菌常见于极端环境诸如酸性温泉，但有些种类可与真细菌一起出现在更有营养的部位（如牛的肠道)。典型的原核生物基因组传统上被认为是含有不到 10Mb DNA 的单个环状染色体。但是，由于更多的原核生物种类得到详细的特性分析，以及很多基因组结构的发现，这种观点最近已受到挑战。例如，已证实原核生物拥有多个环状或线状染色体或混合环状和线状染色体，基因组大小达 30Mb（如 *Bacillus megaterium*)。

真核生物细胞 (eukaryote cell) 被认为最早出现于 15 亿年前。它们有着比其原核同伴复杂得多的结构，具有内膜和包括细胞核的膜边界的**细胞器** (organelle) （框 3.1)。真核生物只有**真核生物** (eukarya) 一个界，但同时包括单细胞物种（例如酵母、原生生物和藻类）和多细胞物种（特别是动物和植物)。在所有已知真核生物中，基因组都由包含在核内的两个或更多的线状染色体组成。每条染色体是一个单独的、很长的 DNA 分子，与组蛋白和其他蛋白质以一种精确的和高度组织的方式组装在一起。染色体的数目和 DNA 含量在不同的物种间变化很大，尽管基因组大小与生物体复杂性具有相平行的趋势，但也有特别的例外（节 3.1.4)。除了主要的细胞器，细胞质和核质中的可溶性部分也都有高度的组织。在核中，已证实多种亚核结构按核基质的情况而安排（框 3.1)。在细胞质中，有一个不同类型蛋白丝的复杂网络，统称为**细胞骨架** (cytoskeleton)，它决定细胞形态和细胞运动，并提供细胞内运输的框架（框 3.2)。

框 3.1 动物细胞的内部结构

细胞器的多样性显示了真核细胞的复杂性，但是由于功能的特化，同样的细胞器不是必须存在于所有的细胞类型中。一些人类细胞表现出极端的变异（例如成熟红细胞缺少细胞核，而终末分化的角质形成细胞根本没有细胞器）。另外的情况，特殊的细胞器存在于极少类型的细胞中，是特定功能所必需的。例如，排列于呼吸道的上皮细胞含有纤毛以将颗粒污染物排出肺部，而成熟精子是唯一拥有鞭毛的人类细胞。但是，大多数细胞含有一套“标准装”细胞器，其功能将在下面讨论。

细胞核：遗传物质的储藏室

细胞核（nucleus）含有一个动物细胞的绝大多数（通常为 99.5%）DNA，以线状染色体形式存在。它被**核被膜**（nuclear envelope）环绕，核被膜由两层被一狭窄腔隙分开并且与内质网（见下文）相连的膜组成。细胞其余的部分称为**细胞质**（cytoplasm），由各种细胞器、膜和称之为胞质溶胶的水性区室组成。细胞核和细胞质之间的交流通过**核孔**（cytosol）进行，核孔开于核被膜中，被核孔复合体围绕，后者是作为核和细胞质之间特定的大分子转运的蛋白质复合体。

在细胞核中，染色体以一种高度有序的方式安排，这反映了存在一种称之为**核基质**（nuclear matrix）的复杂亚结构。核基质是一种蛋白质和 RNA 的网络，染色体通过称为**基质附着区**（matrix attachment region，MAR）的 DNA 序列与之附着。核内还有其他可辨别的结构，这说明不同生化过程有着复杂的组织。**核仁**（nucleolus）是一个独立的区域，**核糖体 RNA**（ribosomal RNA，rRNA）在此合成和加工。rRNA 基因位于此细胞器上，在人类细胞中主要成簇分布在 13、14、15、21 和 22 号染色体的短臂。高 RNA 含量使核仁在电子显微镜下呈现致密的外观。其他核的小体已经被鉴定，但了解得还很少。例如，**Cajal 小体**（Cajal body）[另外也称作**卷曲小体**（coiled body）或宝石] 被认为是**小核核糖核蛋白**（small nuclear ribonucleoprotein，snRNP）颗粒的合成地点，而且可能与基因调节有关。若干转录因子和细胞周期调节蛋白显示位于细胞核内的中央，通常与特定的遗传基因座相关。这说明染色质是在细胞核的特定部位提供给**转录工厂**（transcription factory），而不是传统观念认为的转录因子在整个细胞核内自由扩散。一个特别有趣的例子是 PML 小体，它呈圆环状，主要由早幼粒白血病（premyelocytic leukemia，PML）蛋白构成。在急性早幼粒白血病患者中，一个涉及 PML 基因和另一个编码视黄酸受体的基因的易位产生一种融合蛋白，它不能以正常方式形成 PML 小体。大概是通过允许蛋白正确定位，用视黄酸治疗可恢复细胞形成 PML 体的能力，也可使癌消退。其他核小体似乎与从转录下游过程有关。包括**染色质周纤维**（perichromatin fibril）（新生 RNA 聚积的部位）、**切割小体**（cleavage body）（多聚腺苷酸化和切割的部位）和**染色质间颗粒簇**（interchromatin granule cluster）（与 RNA 剪接有关）。

线粒体：需氧真核细胞的能源库

大多数真核细胞胞质中有一个显著特征，即**线粒体**（mitochondria）是专门产生能量的细胞器。线粒体**呼吸链**（respiratory chain）是一系列五种膜包裹的蛋白复合体，包括各种醌、细胞色素和铁硫蛋白。这些复合体共同负责**氧化磷酸化**（oxidative phosphorylation），在这一反应中，组织的营养物质被分子氧化，新产生的化学能用于生成 ATP。通过随后扩散至细胞的所有部位，ATP 可以供给它储存的能量（通过 ATP 水解为 ADP），所释放的能量用于驱动众多的细胞功能。

尽管大小可变，线粒体直径通常约 1μm，与细菌细胞相似。线粒体有两层膜：相对光滑的外膜和复杂的**线粒体内膜**（inner mitochondrial membrane），后者由于有许多皱褶（嵴）因而有很大的表面积。其内部区室，即**线粒体基质**（mitochondrial matrix），是一种很浓的与能量代谢相关的许多酶和化学中间物的水溶液。

线粒体（以及植物细胞的叶绿体）是仅有的其他含有 DNA 的真核细胞器。此外，它们还有它们自己的核糖体，它不同于在胞质溶胶中发现的核糖体。线粒体核糖体用于翻译从线粒体 DNA 转

框 3.1 动物细胞的内部结构（续）

录的 mRNA。这和其他证据表明线粒体是与真核细胞前体共生的需氧细菌的后代（节 12.2.1）。但是，线粒体 DNA 只能特化极少的线粒体功能，而大多数线粒体蛋白质由核基因编码。对于后者，mRNA 由胞质溶胶中的核糖体翻译，然后生成的蛋白质被转运入线粒体。

内质网：蛋白质和脂质合成的主要部位

内质网由扁平的、单层膜的囊泡组成，囊泡的内部区室即潴泡相互连接、形成贯穿于整个细胞质的通道。内质网从生理上和功能上可以分为两种组分。**粗面内质网**（rough endoplasmic reticulum）被如此命名是因为表面镶嵌着比线粒体中更大的核糖体，而**滑面内质网**（smooth endoplasmic reticulum）缺少任何附着的核糖体。附着于粗面内质网的核糖体所合成的蛋白质穿过内质网膜进入潴泡腔。从这里，它们被转运至细胞外围，在那里它们可以掺入质膜、重新回到内质网或从细胞被分泌。许多人体蛋白是糖基化（有糖基加入它们）的，而这一过程始于内质网。

高尔基体：分泌机器

高尔基复合体（Golgi complex）由扁平的、单层膜的、通常是堆叠在一起的囊泡组成。高尔基体的基本功能是分泌细胞产物（如蛋白质）到外部，以及帮助形成质膜和溶酶体膜。小泡通过一个修剪过程在外围产生，在一些小泡内分泌产物被浓缩了（**分泌泡**，secretory vacuole）。从内质网来的糖蛋白在高尔基复合体被进一步修饰。

过氧化物酶体（微体）：危险化学物专家

过氧化物酶体（peroxisome）（**微体**，microbody）是微小的单层膜囊泡，含有多种酶，可用分子氧氧化它们的底物并产生过氧化氢。过氧化氢被过氧化物酶体中一种主要的酶——触酶——用于氧化大范围的化合物。活性氧类，如过氧化物和超氧化物基团，对细胞有很高的毒性，必须仔细包容。

溶酶体：细胞内消化器官

溶酶体（lysosome）是单层膜包裹的小囊泡，含有水解酶如核糖核酸酶和磷酸酶。溶酶体的功能是消化通过吞噬或胞饮作用进入细胞的物质，以及帮助降解随细胞死亡产生的细胞组分。

质膜（细胞膜）：细胞的防卫前沿

质膜（plasma membrane）由双层磷脂组成，其中疏水脂质基团位于内侧，它们被亲水的磷酸基团夹在中间。磷酸基团两侧与细胞质和细胞外的水性环境接触。除了提供一般的保护屏障，质膜还有许多重要的功能：

- 它具有选择通透性，调节许多离子和小分子进出细胞。它含有活性转运系统，针对各种离子如 Na^+、K^+ 和 Ca^{2+}；各种营养物质如葡萄糖和氨基酸以及许多重要的酶；
- 它含有许多完整的膜蛋白或锚定于膜一侧的蛋白，它们在细胞-细胞信号传递中起着重要的作用。

胞质溶胶：一种高浓度和结构性的水溶液

胞质溶胶（cytosol）是细胞质的水性成分，占大约一半的细胞体积，是包括大多数蛋白合成和中间代谢在内的主要代谢活动的部位。除了含有可溶性组分，胞质溶胶被统称为**细胞骨架**（cytoskeleton）的一连串的蛋白丝进行了非常高度地组织。细胞骨架在细胞运动、细胞形态和胞内运输中起着重要作用。

纤毛和鞭毛：运动装置和混合器

纤毛和鞭毛是从质膜延伸出的结构，用于帮助运动。**纤毛**（cilia）是能前后摆动或旋转的微小结构。单细胞生物通常依靠成千上万纤毛的协同摆动获得自动力，但是这些结构也存在于动物的固定细胞，它们在那里用于移动液流。在人体，纤毛见于沿呼吸道排列的上皮细胞（此处它们驱动那里包含的黏液和任何颗粒状物质移出肺部）和输卵管（这里它们提供动力使卵朝子宫移动）。纤

框 3.1 动物细胞的内部结构（续）

毛也存在于一种称为**节**（notle）（见 3.7.5 节）的微小胚胎结构，在那里它们的转动引起节周液体移向胚胎的另一侧。这一过程被认为对胚胎的左-右轴特化非常重要（框 3.7）。**鞭毛**（flagella）是以鞭子样方式运动的较大的结构。很多单细胞生物通过鞭毛获得动力，但与纤毛不同，每个细胞通常只有一条或两条鞭毛。唯一拥有鞭毛的人体细胞是精子，它们用这种细胞器作为获得推进力的一种方式。纤毛和鞭毛都是由成束的特征性“9＋2”（外部 9 组成对微管围绕中心两个单独的微管）或“9＋0”结构的**微管纤丝**（microtubule filament）组成。这些结构附着在质膜下的**基体**（basal body）上，基体由细胞分化期间重复重制的中心粒生成（框 2.1）。运动由外围的双层微管相互滑动产生，由**动力蛋白**（dynein）控制。在人类，动力蛋白缺乏会导致以反复发生的呼吸道感染、偏侧性缺陷为特征的**纤毛无力综合征**（immotile cilia syndrome），还会导致不育，原因是卵子不能到达子宫以及精子不能摇动它们的尾部。

框 3.2 细胞骨架：细胞运动和细胞形态的关键以及细胞内运输的主要框架

真核细胞的细胞骨架是一个**蛋白纤丝**（protein filament）的内部框架。蛋白纤丝提供稳定性，产生运动和改变细胞形态所需的力，帮助细胞器进行细胞内转运并允许细胞与周围环境之间通讯。有三种类型细胞骨架纤丝：肌动蛋白微丝、微管（由微管蛋白构成）以及中间纤维（由各种不同的蛋白组成，其中一些是细胞类型特异性的）。肌动蛋白和微管在进化中是高度保守的。微丝和微管是非常动态的结构，由于构成肌动蛋白和微管蛋白的亚单位在多聚化过程中的不对称性，它们也是极化的。所以，聚合体在生成时，新的亚单位同时在两端但是以不同的速率附着，导致出现一个称作**正（＋）末端**［plus（＋）end］的快增长末端和一个称作**负（－）末端**［minus（－）end］的慢增长末端。特殊类型的动力蛋白能够与极化的纤丝结合，并且依靠 ATP 水解供能，以一个方向沿着它们稳定地移动。

肌动蛋白丝（actin filament）（也称为**微丝**，microfilament）能够以平行束状或网格状排列。它们以应力纤维的形式提供机械支撑，允许有控制的细胞形态改变（例如顶点压缩，细胞分裂过程中的压缩），并且通过形成诸如丝足、薄足（延伸至细胞，允许它沿表面爬行）等结构帮助细胞运动。肌动蛋白丝是诸如微绒毛、黏着斑以及精子头部的顶体等特殊结构的基础。它们由**肌动蛋白**（actin）的双链螺旋多聚体组成，直径大约 7～8nm。它们与动力蛋白的**肌球蛋白超家族**（myosin superfamily）成员相互作用。在肌肉细胞中，肌动蛋白-肌球蛋白的相互作用形成收缩单位，提供给肌肉细胞它们的收缩力。在非肌肉细胞，肌动蛋白还有更一般的细胞功能，例如允许膜内陷，如**胞吞作用**（endocytosis）。

微管（microtubule）是直径为 25～30nm 的长而中空的圆柱体，比肌动蛋白丝刚硬得多。它们由一系列基于交错的 **α 微管蛋白**和 **β 微管蛋白**（β-tubulin）残基的多聚体装配组成。与微管有关的是发动蛋白的两个超家族的成员，即 **α 动力蛋白**（α-tubulin）和**驱动蛋白**（kinesin），它们负责沿细胞内的微管轨道移动特殊的“货物”。例如，这就是线粒体、高尔基（dynein）体堆、分泌泡如何到达细胞内它们合适位置的方式。动力蛋白和驱动蛋白以相反的方向沿微管移动。

典型的情况是，动物细胞中每个微管的负末端聚集在位于接近核的胞质部分的**微管组织中心**（microtubule-organizing centre，MTOC）。一个单独的、明确定义的这种类型的中心称为**中心体**（centrosome），可见于大多数动物细胞。在细胞分裂时，微管形成**有丝分裂纺锤体**（mitotic spindle）和微管纤丝，附着于位于被复制染色体的着丝粒的蛋白复合体（**动粒**，kinetochore），确保染色体以一种有序的方式移入两个子细胞（框 2.1）。微管也形成**纤毛**（cilia）和**鞭毛**（flagella）的核心，这已在框 3.1 中讨论。

框 3.2　细胞骨架：细胞运动和细胞形态的关键以及细胞内运输的主要框架（续）

中间纤维（intermediate filament）直径为 7～11nm，具有主要的结构功能。它们没有极化，没有相关的动力蛋白。神经丝（神经细胞特有）和角蛋白（上皮细胞特有）是中间纤维蛋白。

3.1.2　细胞大小和形态可以有巨大的变化，但扩散率有确定的上限

有很多因素影响细胞大小（Su and O'Farrel，1998；Saucedo and Edgar，2002）。细胞依靠扩散来协调它们的代谢活动、它们与环境和其他细胞的相互作用。当它们长大时，表面积与体积比例降低。有人认为原核细胞简单的内部结构限制了他们的最大体积——典型的细菌细胞的直径为 1μm。真核细胞复杂的内膜和区室化可能对于允许它们长得更大很重要。但是，代谢活跃的内部区域距离细胞表面很少超过 15～25μm，这就限制了细胞的大小大约为 50μm。事实上，多细胞生物的细胞平均直径在 10～30μm 的范围。一些特化的细胞能够长到比这大得多。例如神经元，其长度能够达到 1m，尽管这只反映其长而很细的轴突的凸起，而细胞体很好地保持在上面讨论的参数内。卵细胞也是很大的细胞。哺乳动物卵细胞的直径大约为 100μm，但是其他用于储存发育所需营养的动物卵细胞可以大得多。已知最大的卵细胞是鸵鸟卵，可达 20cm 长，体积相当于 24 个鸡蛋。

3.1.3　在多细胞生物中，体细胞和生殖系之间有根本区别

在多细胞生物中，生长和繁殖功能是分开的。一个特化的**生殖细胞**（germ cell）群体分出来执行繁殖功能，而剩下的**体细胞**（somatic cell）的作用是提供携带这些生殖细胞的容器和完成繁殖所需的手段（Wylie，2000）。在进化学的术语中，动物可以被认为是允许物种繁衍的精子和卵子的孵育器和促进器。在植物和原始动物中，普通体细胞在生物体的整个生命中都可以成为生殖细胞。但是，在大多数我们详细了解的动物（昆虫、线虫和脊椎动物）中，生殖细胞在发育的很早的阶段就分离出来作为专门的**生殖系**（germ line）并且成为配子的唯一来源。生殖细胞是机体中能进行减数分裂的唯一细胞，所以也是唯一能产生参加受精的单倍体子代的细胞。哺乳动物的生殖系细胞起源于**原始生殖细胞**（primordial germ cell，PGC），就人类而言 PGCs 出现于发育的第二周。

3.1.4　在多细胞生物中，没有两个细胞携带完全相同的 DNA 序列

如 2.1 节中所述，细胞的参考 DNA 含量，C 值，由单倍体染色体组（n）决定，如精子和卵细胞。这个值在不同的生物中变化很大，而 C 值与生物复杂性之间直接关系的缺乏称为 **C 值颠倒**（C value paradox）。例如，尽管大多数哺乳动物的单倍体基因组大小范围约为 2500～3500Mb DNA，当发现一个人体细胞的单倍体 DNA 含量只有洋葱的 19%、一些百合的 4%并且更显著的只有单细胞真核生物 dubia 阿米巴的 0.5%时，实在让人有点吃惊（见表 3.1 及基因组大小数据库 http://www.cbs.dtu.dk/databases/DOGS）。

表 3.1　C 值颠倒：基因组大小不是简单地与生物体复杂性相关

生物体	基因大小	基因数
单细胞		
大肠埃希菌	4.2Mb	4300
酿酒酵母	13Mb	6300
dubia 阿米巴	670 000Mb	?
多细胞		
秀丽新小杆线虫	95Mb	约 20 000
黑腹果蝇	165Mb	约 13 000
洋葱	15 000Mb	?
小鼠	3000Mb	约 30 000
人类	3300Mb	约 30 000

由于染色体组（**倍性**，ploidy）的拷贝数不同，在单一个体中，DNA 含量也有相当大的变异。虽然大多数人类体细胞是双倍体，也有例外。如框 3.1 中讨论的，一些终末分化细胞，如红细胞、角质形成细胞和血小板，没有细胞核，被称为**无倍体**（nulliploid）。由于没有细胞分裂的 DNA 复制（**核内有丝分裂**，endomitosis）或细胞融合，其他一些细胞是**多倍体**（polyploid）。例如，肝细胞的倍性范围从 $2n$～$8n$，心肌细胞（心脏肌肉细胞）从 $4n$～$8n$，而骨髓的巨型巨核细胞倍数从 $16n$～$64n$。后者每个细胞产生数千个无倍体血小板细胞（图 3.3）。细胞融合产生的多倍体导致一些自然生成细胞（如肌纤维）具有多个双倍体核，形成**合胞体**（syncytium）（图 3.3）。

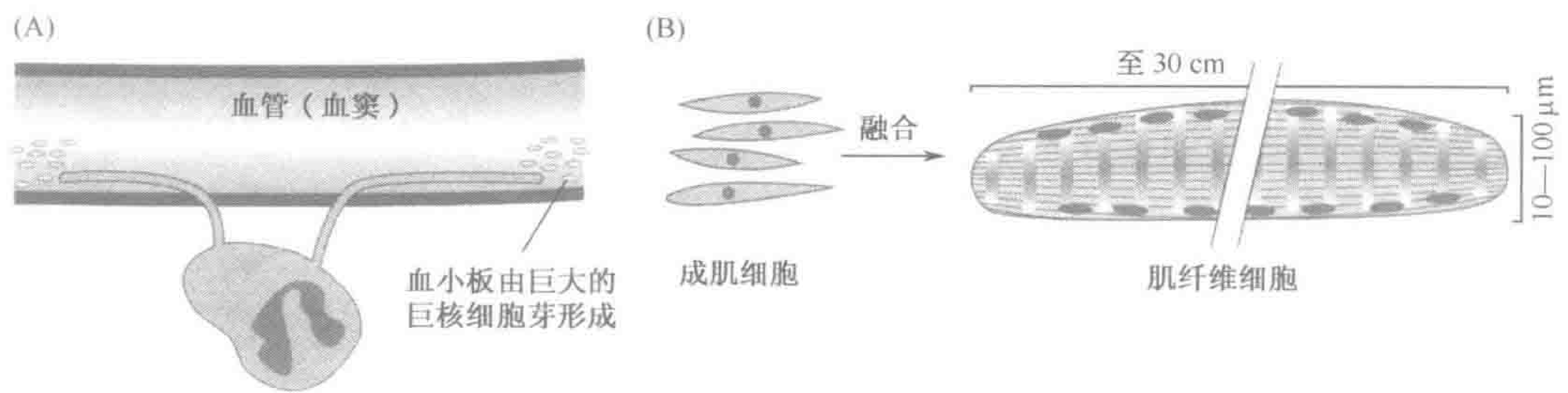


图 3.3　一些细胞由其他细胞破碎或融合形成

血小板由巨大的巨核细胞芽生形成，它们无细胞核。肌细胞由大量的成肌细胞融合形成。

由于 DNA 序列的遗传差异（突变），来自不同生物的细胞 DNA 可以有相当大的差异。序列的差异程度大致与两个物种相距的进化距离成比例。所以，人类细胞的 DNA 在序列上与黑猩猩细胞密切相关（98%～99%），但分别与小鼠、蛙和果蝇等细胞相比则趋异越来越大。同一物种中不同个体之间细胞 DNA 的差别也可显示出突变的差异。当比较两个无关人的 DNA 时，大约每 1000 个核苷酸有一次改变。单个个体的细胞 DNA 序列之间也有差异。这些差异可以来自三种途径。

▶ **特殊细胞中的程序差异**。包括精细胞的例子，它具有一条 X 或一条 Y 染色体，因而

携带不同的性染色体 DNA 组合；还有成熟的 B 和 T 淋巴细胞，它们经历了细胞特异性 DNA 重排，导致产生的抗体或 T 细胞受体有细胞与细胞的差异。

- **随机突变和 DNA 不稳定性**。所有细胞的 DNA 都不断地受到环境损伤、化学降解和修复，这会影响 DNA 复制的精确性。这种小而重要的错误率意味着每一轮细胞分裂都有 DNA 序列新的突变发生。所以，在发育过程中每个细胞都会建立起独特的突变谱，在同一个体中没有两个细胞会有精确相同的 DNA 序列。这些突变大多数根本没有表型效应，而且，即使发生于必需基因的突变当其发生于单独的细胞时，对整个个体也没有影响（除非它们引起细胞增殖；17 章）。只有当缺失突变发生于生殖系、干细胞或发育早期的体细胞时，影响才可能见于表型的水平。
- **嵌合性和克隆化**。在极偶然的情况下，一个个体可能会显示具有两个或更多的携有非常不同 DNA 序列的细胞克隆，这反映了发育早期异卵双生胚胎的融合或一个胚胎中来自另一个胚胎的细胞的克隆化（图 4.10）。在表层，这样的差异也存在于接受了移植或输血的个体。

3.1.5　多细胞生物的细胞可以通过原位或培养研究

单细胞生物通常可以在相当它们正常环境的培养基上培养来研究。对于多细胞生物而言，单个细胞的环境很难复制。所以，多细胞生物的细胞既可以在完整生物体的自然构造下研究，也可在它外部的人工环境下研究。有几种不同的方法可以使用。

- **原位**（*in situ*）（作为来源动物的一部分）分析。对于最简单的动物，例如秀丽新小杆线虫，有可能通过显微镜在整个活体标本中研究单个细胞。使用绿色荧光蛋白作为活体标记，可以允许在较大动物的活体标本进行原位研究，只要被研究的组织是光学半透明的。许多动物（如果蝇、斑马鱼、小鼠）的胚胎在早期是半透明的。对不透明的标本，可能必须在处理后制备组织切片。
- 作为**组织移植物**（tissue explant）。通过单独或联合研究特定组织移植物的行为，许多发育生物学研究就方便了。这用来检测特定发育分子的作用或研究一个细胞群体对另一个细胞群体的诱导作用（节 3.4.2）。
- 作为**原代细胞**（primary cell）。组织移植物可以被分解，使单个细胞分离并单独在培养基上生长。这不总是一种直截了当的方法，因为一些细胞类型会很难生长，而且维持原代细胞的代价会相当大。
- 作为**稳定的细胞系**（cell line）。一个稳定的细胞系通常很容易在培养基中生长所以会很容易扩增并运送给全世界的研究者。由于这些细胞可以鉴定出种种诊断特征，所以研究者可以自信，他们从一个特定的细胞系获得的结果，能与在别处使用同一细胞系的其他研究结果参考或结合。

用组织移植物或培养的原代细胞研究有很多优点，但它们难以操作而且是非永久性的资源。这是因为大多数动物细胞在培养中将经历一定次数的分裂，然后**衰老**（senescence），此时它们将停止分裂并退出细胞周期，进入称作 G_0 的静止期（Campisi，1996）。分裂的次数依赖于细胞类型、物种和来源生物的年龄。例如，人胎儿成纤维细胞在培养中可分裂大约 60 次，而取自成人的成纤维细胞则会经历大约这个数目一半的分裂次数。分裂潜能与物种的寿命之间似乎有某种联系。例如，人胎儿成纤维细胞达到

的分裂次数（大约 60 次分裂，最长寿命 120 年）介于胎鼠成纤维细胞（大约 15 次分裂，最长寿命 3 年）和 Galapagos 巨龟胚胎成纤维细胞（大约 125 次分裂，最长寿命 175 年）之间。

原代细胞的局限性可以通过建立**永生细胞系**（permanent cell line）来克服，它可以在培养基中无限分裂。用肿瘤切除物已建立了许多细胞系，它们已经失去了某种生长限制。例如，广泛使用的 HeLa 细胞系来源于 1955 年从一位名叫 Henrietta Lacks 的患者身上切除的宫颈肿瘤。肿瘤组织来源的细胞系的问题是，与原始的来源组织相比，它们通常具有不同的核型，这使得随着传代次数的增加，表型差异变得越来越显著。细胞系也能通过诱导原代细胞在培养中经历生长转化（永生化）获得。这一过程与见于肿瘤中的失去生长限制类似，可由于获得突变自发产生，或可通过辐射、化学致突变剂或转化病毒诱导产生。最近，已经通过用特定的致癌 DNA 序列转染原代细胞建立了细胞系。数千种细胞系，代表不同的人类和动物组织，保存在诸如美国培养物类型收藏库（ATCC）、欧洲细胞培养物收藏库（ECACC）以及国际实验室细胞系收藏库（ICLC）等细胞库中（表 3.2；Stacey and Doyle，2000）。

表 3.2 人细胞系及其用途举例

ATCC	编码细胞系	来源及应用
淋巴细胞系		
CCL213	DAUDI	B 细胞系，来源于 Burkitt's 淋巴瘤，用于癌研究
CRL1432	NAMALWA	B 细胞系，来源于 Burkitt's 淋巴瘤，用于生产干扰素
TIB152	JURKAT E61	T 细胞系，来源于急性成淋巴细胞白血病，产生大量的白介素-2
CRL1942	SUP-T1	T 细胞系，来源于 T-成淋巴细胞白血病，支持 HIV 复制
骨髓细胞系		
CCL240	HL-60	来源于早幼粒细胞白血病细胞，用于髓系谱系分化研究
TIB202	THP	来源于急性单核细胞白血病，用于研究巨噬细胞的激活和成熟
肾细胞系		
CRL1573	293	腺病毒转化的胎肾细胞，用于研究腺病毒
CRL2190	HK-2	人乳头状瘤病毒 E6/E7 基因转化的近曲小管细胞系，用作近曲小管研究的模型系统
肝细胞系		
HB8064	Hep3B	产生许多血浆蛋白
HTB52	SK-HEP-1	来源于肝腺癌，诱导裸鼠肿瘤
卵巢细胞系		
HTB75	Caov-3	来自卵巢腺癌，产生 p53 突变，用于细胞因子和癌研究
CRL1572	PA-1	来自卵巢畸胎瘤，用于发育研究和药物检测

细胞系的一个重要应用是维持从特殊的人类患者获得的可更新的遗传资源。这需要若干程序从组织配型实验室质量监控到连锁分析和基于标记的制图。在这种情况下，细胞的表型不重要，但基因型必须精确地保留。要做到这一点，最直接的方法是通过用

Epstein-Barr 病毒（EBV）转化制作类淋巴母细胞细胞系。EBV 保持游离状态（非整合的），所以不会改变内源基因组。人类 B 细胞有 EBV 受体，而且一旦感染，它们有很高的成功率变得永生化，产生可在培养基中无限增殖（或低温保存以备将来使用）的细胞系。

3.2 细胞的相互作用

3.2.1 细胞间的通讯涉及特异受体对信号分子的感知

在多细胞生物中，单个细胞的存活和繁殖从属于生物作为一个整体的存活和生殖。为了生物体的利益，体细胞必须相互协作。由于这个原因，多细胞生物内的细胞需要通讯以协同和调节生理和生化功能。细胞通讯的基础是一个细胞群体产生的**信号分子**（signaling molecule）以及通常见于效应细胞表面的**受体**（receptor）对它们的识别（表 3.3）。动物细胞肯定都布满信号分子受体，而且这些受体与内部的**信号转导通路**（signal transduction pathway）连接，这样可以对转录因子活性进行调节，最终改变基因表达模式（见 Twyman 2001 的综述）。

表 3.3 重要的信号分子及其受体类型

	受 体	举 例
分泌信号		
多种（肽、蛋白、有机小分子、物理刺激）	G 蛋白偶联受体 单链多肽。中央亲水区形成一个七次跨膜区，N 端区与配体结合，内部 C 端区与鸟嘌呤核苷酸结合蛋白联合	肾上腺素、5-羟色胺、胰高血糖素、及神经激肽的受体。也包括嗅觉受体、味觉受体及视觉受体视紫质
生长因子，ephrins	受体酪氨酸激酶 二聚体，N 端配体结合区，内部激酶区，跨膜一次	胰岛素、成纤维生长因子、表皮生长因子及 ephrins 受体
细胞因子	酪氨酸激酶活性相关受体 寡聚体，N 端配体结合区，内部 Janus 激酶（JAK）联合区	生长激素、促乳素、促红细胞生成素、集落刺激因子、白介素、干扰素受体
转化生长因子-β 家族	丝氨酸/苏氨酸激酶活性相关受体 寡聚体，N 端配体结合区，内部 SMAD 蛋白联合区	TGF-β 蛋白、骨形态发生蛋白、Nodal、神经胶质来源的嗜神经因子受体
Hedgehog 家族	膜片	哺乳动物发育过程中 SHH 和印度 SHH 受体
Wnt 家族	卷曲片	哺乳动物发育过程中 Wnt 家族蛋白受体
类固醇激素，视黄醛衍生物	核受体 二聚体，位于胞质或核，转换为转录因子与配体联结	类固醇激素、甲状腺激素、维生素 D、维甲酸受体
固定信号		
Delta/Serrate 家族	切迹	神经发育过程中的 Delta 和 Serrate

细胞间信号可以在不同的范围产生。大多数人熟悉**内分泌信号**（endocrine signaling）的概念，激素从身体一部分的内分泌腺释放，通过血流运动抵达远处的靶细胞群。内分泌信号对保持体内平衡非常重要，而且在血管系统建立后的后期发育中起着主要作用。然而在早期发育过程中，最重要的信号只跨越较短的距离。**旁分泌信号**（paracrine signaling）包括信号分子从一个细胞群的释放以及这些分子通过较短的距离扩散至效应细胞。**邻分泌信号**（juxtacrine signaling）是基于邻近细胞的相互作用，而且反映了这样一个事实：一个细胞产生的信号保持着与质膜的联系。细胞也可对存在于它们周围环境即**细胞外基质**（extracellular matrix）的分子作出反应（节 3.2.4）。

3.2.2 激活受体引发可能包括酶级联反应或第二信使的信号传导通路，导致转录因子的激活或抑制

一旦特定信号蛋白的受体被激活，一系列的事件会发生，最终导致效应细胞内基因活动模式的改变。这个链条内连接的数目和本质，依赖于所涉及的特定信号分子。类固醇和甲状腺激素以及发育调节蛋白维甲酸可以扩散通过质膜，所以它们的受体位于细胞内。一旦与它们的配体结合，这些受体可直接作为转录因子调节下游基因的表达。所以，类固醇样分子利用单步信号传导通路（Tsai and O'Malley，1994）。

大多数其他信号分子留在细胞外与跨膜受体结合。配体与受体的细胞外区域结合，使其细胞内区域结构发生改变，通常会刺激潜在的酶活性。对于受体激酶，这种结构改变会激活其激酶活性，激酶活性对受体和受体相关蛋白的完整性都是必需的。这首先引起激酶自身磷酸化，进而允许它磷酸化并从而激活细胞内的其他蛋白质。这些靶蛋白通常自身也是激酶，可以继续磷酸化更下游的蛋白质。最终，激酶活性的级联反应到达转录因子，使其被激活或抑制，引起基因表达适当的改变。在某些情况下，信号级联很短（例如细胞因子调节的 JAK-STAT 途径；Pellegrini and Dusanter-Fourt，1997；Hou *et al.*,2002；图 3.4），而其他的则有很多步（例如生长因子调节的 MAP 激酶途径；Robinson and Cobb，1997；图 3.5）。

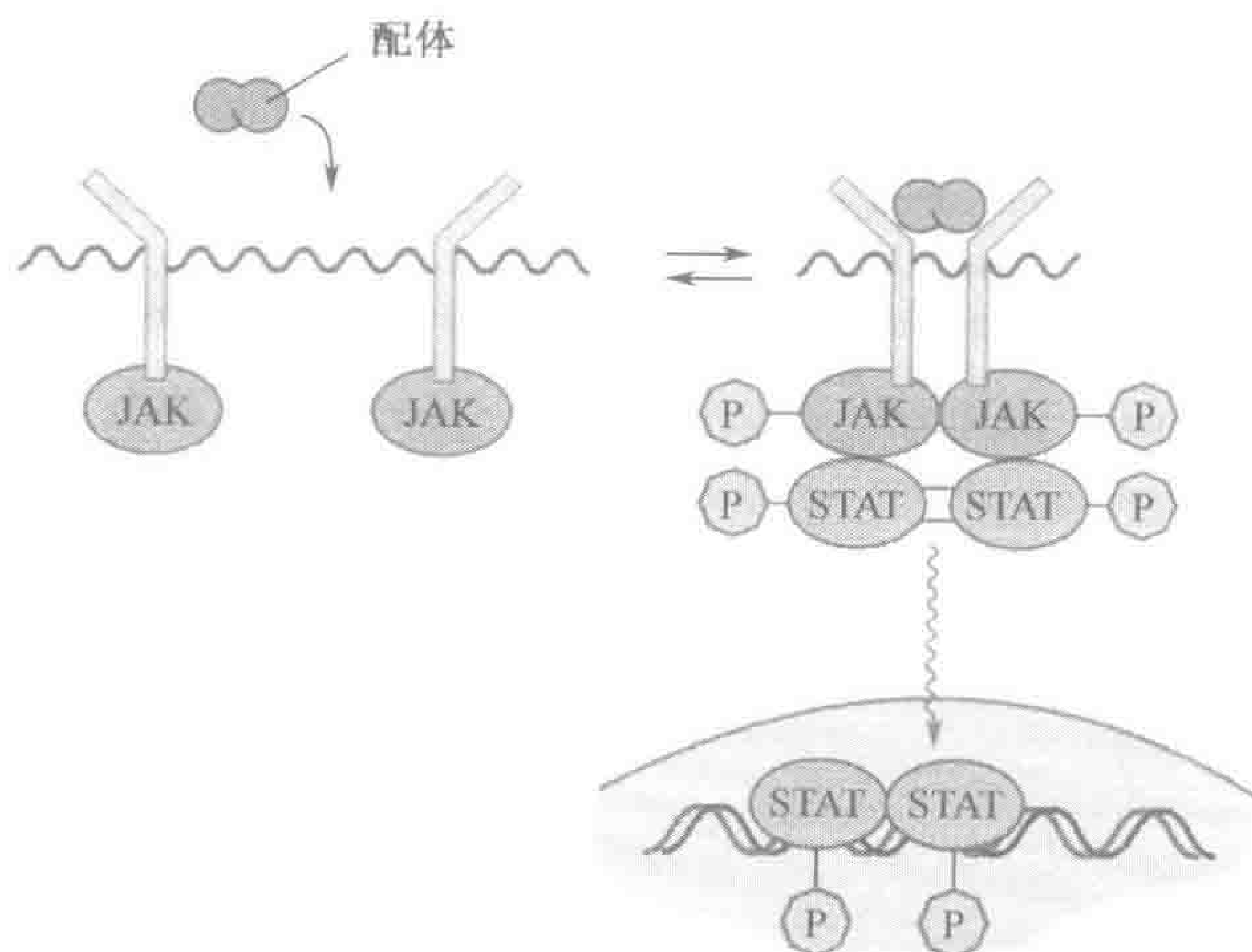


图 3.4 细胞因子受体是二聚体或寡聚体，每条肽链跨膜一次

受体没有内在的酪氨酸激酶活性，但它们在结构上与胞质中的 Janus 家族酪氨酸激酶（Janus 激酶，JAK）联结。配体结合引起受体二聚化，这导致联结的 JAKs 发生相应的自转磷酸化，后者再激活并使受体自身磷酸化。然后受体可以捕获失活的 STAT（转录的信号转导子和激活子的转录因子），它通过其 SH2 结构域与磷酸酪氨酸结合。接着 STAT 被 JAK 磷酸化，使它们二聚化并且转移至核，激活下游基因。重绘自 Twyman（2001）. Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

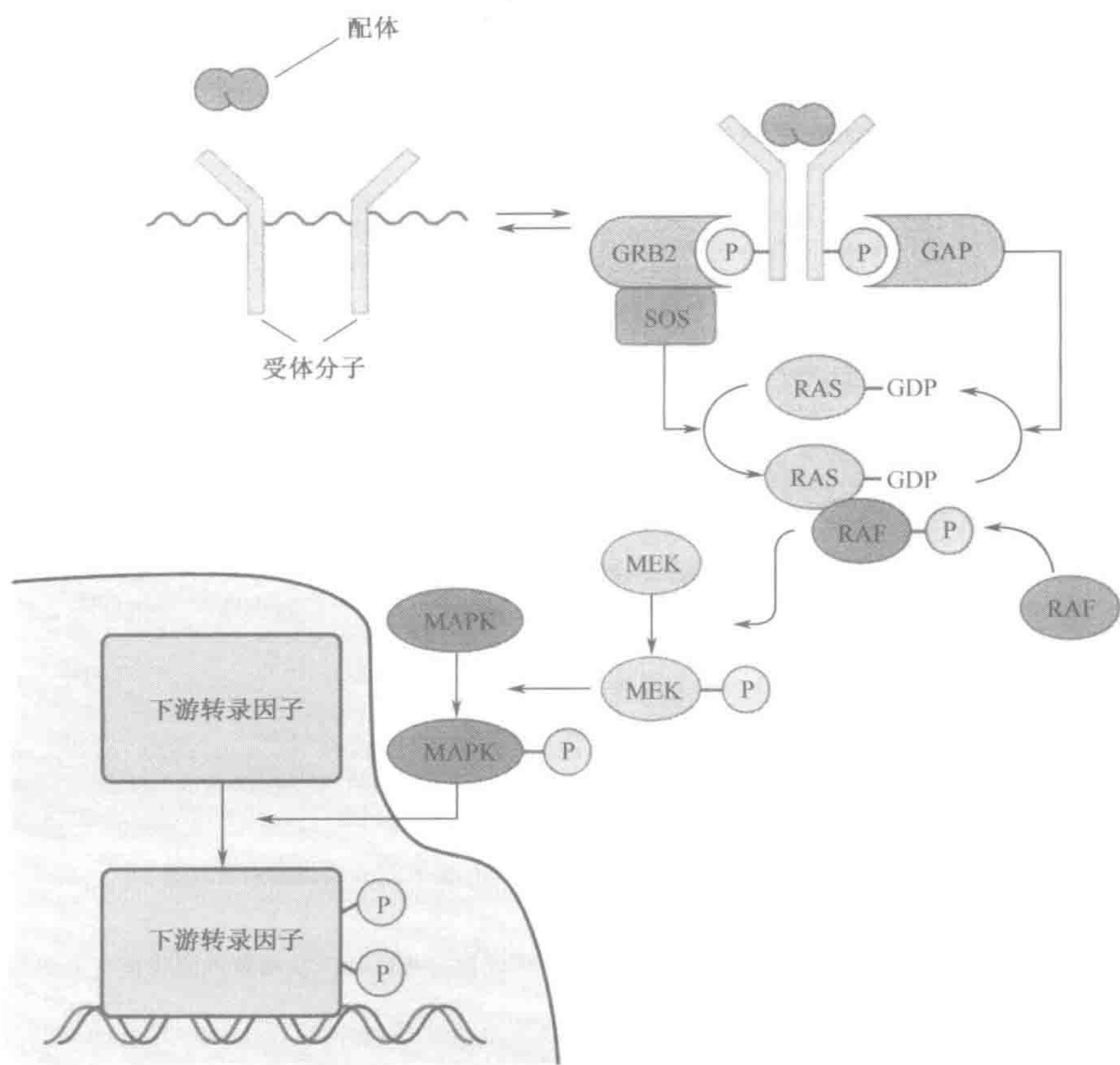


图 3.5　生长因子作为酪氨酸激酶受体的配体

配体结合刺激受体二聚化并激活受体内在的胞质酪氨酸激酶活性。受体不仅能使其他蛋白磷酸化，还能磷酸化自身（自磷酸化），导致特异结合到磷酸酪氨酸残基的蛋白质被俘获，包括调节 Ras 活性的酶。活化的 Ras 将 Raf 招至膜上其激酶活性受刺激处，然后 Raf 激活 MEK，后者再依次激活 MAP 激酶。MEK 和 MAP 都能激活潜在的转录因子，从而改变核内基因表达模式。重绘自 Twyman（2001）. Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

对于 G 蛋白偶联受体，信号传导通过第二信使（小分子如 cAMP、钙或脂质）的激活完成。配体结合引起鸟嘌呤核苷结合蛋白（G 蛋白）与受体连接，将 GDP 转化为 GTP，这使其分解为 α 和 βγ 单位。然后这些单位每一个都可与下游蛋白质相互作用，以调节第二信使水平（Bourne，1997）。依靠 G 蛋白的特殊形式，配体结合可能刺激或抑制腺苷环化酶，引起细胞内 cAMP 水平的改变（Houslay and Milligan，1997）。其他 G 蛋白刺激脂质如肌醇-1，4，5-三磷酸和甘油二酯的产生（Speigel *et al.*，1996）或钙离子的释放（Clapham，1995）。依次的，第二信使激活下游蛋白激酶如 cAMP 依赖的蛋白激酶 A 和钙依赖的蛋白激酶 C，后者继续磷酸化且进而改变特定转录因子的活性。

在上述信号通路间存在广泛的交叉感知。例如，蛋白激酶 A 和蛋白激酶 C 都与 MAP 激酶通路相互作用，受体酪氨酸激酶可刺激 JAK-STAT 通路及影响第二信使水

平，而细胞因子受体可影响 Ras 活性。特定细胞做出的反应依赖于到达其表面的所有信号的总和，且依赖于存在的特定受体和信号成分。

3.2.3 细胞的组合形成组织需要细胞黏着

细胞通过表达**黏着分子**（adhesion molecule）组合成组织（Cunningham，1995；Gumbiner，1996）。在本质上，黏着分子以与任何其他受体-配体相互作用相同的方式发挥作用，尽管此时受体和配体附着于邻近细胞的表面，而且每个细胞的表面有成千上万个这样的分子。细胞可能通过在表面表达互补的黏着分子直接黏着在一起，且/或它们可与细胞外基质形成联合。在发育期间，黏着分子表达的改变允许细胞相互之间生成和打破连接，这便于细胞单个或成片迁移，引起大规模的重排（McNeil，2000；Irvine and Rauskolb，2001）。在成熟生物体中，细胞间的相互黏着通常通过特化的连接区的形成来加强，称为**细胞连接**（cell junction）（Steinberg and McNutt，1999；Perez-Moreno *et al.*，2003；表 3.4）。

表 3.4 细胞连接

黏着连接	锚定连接，在其中细胞骨架的肌动蛋白纤丝与钙黏着蛋白在细胞表面连接，从而将邻近的细胞连接为一个连续的黏着带
桥粒	锚定连接，在其中细胞骨架的中间纤维与钙黏着蛋白在细胞表面连接
黏着斑	锚定连接，在其中细胞骨架的肌动蛋白纤丝与整联蛋白在细胞表面连接，为细胞外基质提供点状的附着位点
半桥粒	锚定连接，在其中细胞骨架的中间纤维与钙黏着蛋白在细胞表面连接，用于将上皮细胞与基片连接
间隙连接	它们是连接邻近细胞胞质溶胶的小孔。它们由六个相同的连接蛋白组成，在每个膜上形成一个通道。连接蛋白复合体附着在邻近的细胞上形成间隙连接
紧密连接	相互连接的跨膜蛋白将上皮细胞封入连续的片层。依靠连接中蛋白的密度，屏障可能对特定大小的分子完全不能通过或选择性通过

有四类主要的细胞黏着分子（Humphries and Newham，1998；Hynes，2002）。

- ▶ **钙黏着蛋白**（cadherin）。它们是钙依赖的跨膜黏附蛋白，识别和结合其他细胞表面相似的钙黏着蛋白。相似的分子结合的过程称为**同嗜性结合**（homophilic binding）。在哺乳动物中存在大约 30 种钙黏着蛋白，其中最重要的是 E-钙黏着蛋白（主要局限于上皮细胞）和 N-钙黏着蛋白（主要在神经系统表达）。
- ▶ Ig-CAMs。它们是结构与免疫球蛋白（Ig）家族相似的钙依赖的跨膜黏附分子。例如，神经细胞黏附分子（NCAM）表达于神经系统和发育中的体节。
- ▶ **整联蛋白**（integrin）。它们是钙依赖的跨膜黏附分子，通常介导细胞-基质相互作用（见下文），但是某些白细胞整联蛋白也与细胞-细胞黏着有关。
- ▶ **选择蛋白**（selectin）。此类分子在炎症反应中由内皮细胞表达。它们识别嗜中性表面的糖基并引导这些细胞到炎症位点。

3.2.4　胞外基质为体内所有组织提供支架，并且也是控制细胞行为的信号的重要来源

机体细胞之间的空间并不是空的。细胞外空间填充着三维排列的、包埋在称为**糖胺聚糖**（glycosaminoglycan）的复合糖凝胶中的蛋白纤维。这种物质，即**胞外物质**（extracellular material，ECM），成分是可变的，它依赖于此处的细胞向其周围环境分泌何种物质。细胞也可通过分泌修饰酶，如蛋白酶，来影响 ECM 的结构，而这会影响细胞的行为（Streuli，1999）。依次的，ECM 提供结构支架，对维持组织完整性起着主要的作用，而且也可通过与跨膜受体作用在发育过程中引导细胞的行为（Giancotti and Ruoslahti，1999；Bokel and Brown，2002）。

ECM 的成分可分为七组（表 3.5）：

表 3.5　细胞外基质的分子成分

成分	分布	结构/功能
胶原	多种组织	胶原是大的糖蛋白三聚体。有几种化学性质截然不同的胶原形式，它们在不同的组织中含量各不相同。含量最丰富的形式是Ⅰ、Ⅱ和Ⅲ型胶原，它们易于通过在赖氨酸残基之间交联稳定地形成纤维。它们通过与整联蛋白相互作用提供结构支持及影响细胞的分化和迁移。反过来，Ⅳ型胶原更易形成格子而非纤维，它是基底层的重要成分，通过与层粘连蛋白结合间接与细胞相互作用。
纤连蛋白	多种组织	纤连蛋白是糖蛋白的二聚体，其主要功能是帮助细胞-基质黏着。纤连蛋白有胶原和肝素结合区，有助于组织 ECM。纤连蛋白通过整联蛋白与细胞相互作用并影响细胞形态、运动和分化。
层粘连蛋白	上皮组织	层粘连蛋白是三聚体蛋白质，由 A、B1 和 B2 亚单位组成。它们与Ⅳ型胶原相互作用形成基底层，其主要功能是通过与整联蛋白和其他细胞表面分子相互作用帮助细胞黏着到基底层，每种层粘连蛋白亚单位有多种组织特异形式。
巢蛋白	上皮组织	与层粘连蛋白按一定比例连接。含有可能允许其与细胞表面整联蛋白相互作用的序列。
弹性蛋白	多种组织，但在血管、皮肤及其他能伸展和变形的结构中较丰富	弹性蛋白是非糖基化的蛋白，通过交联形成网络和片层。蛋白具有天然的弹性回缩力，为组织提供变形后恢复形态的能力。
肌腱蛋白	胚胎，成人神经组织	肌腱蛋白是六聚体糖蛋白，在细胞迁移控制中起重要作用。根据细胞表面表达的受体不同，它可在一些细胞中产生黏着效应，而在其他细胞中产生排斥效应。
玻连蛋白	血液和其他组织	这种蛋白通常与纤连蛋白连接，尽管不是广泛分布，但可与特殊类型的整联蛋白作用以利于细胞-基质黏附。
蛋白聚糖	多种组织	种类极多的分子家族，由一个核心蛋白，如核心蛋白聚糖或多配体蛋白聚糖和一个或多个糖胺聚糖（如硫酸软骨素、硫酸皮肤素、肝素、硫酸乙酰肝素）组成。通常在 ECM 中采用较高度有序的结构。这些分子调节细胞黏着，也可连接生长因子和其他生物活性性分子。
透明质酸	多种组织	未硫酸化、并且没有与蛋白共价结合形成蛋白聚糖的唯一的糖胺聚糖。帮助细胞迁移，特别是在发育和组织修复期间。

▶ 结构蛋白，如胶原和弹性蛋白；

▶ 黏着蛋白，如层粘连蛋白和纤连蛋白；

▶ 蛋白聚糖，它组成与各种糖胺聚糖有关的核心蛋白。蛋白聚糖不同于糖蛋白之处在于前者糖基占分子总量的95%以上。
▶ 游离糖胺聚糖，透明质酸。

ECM的结构和黏附蛋白在决定其功能方面起主要作用。例如，富含胶原组织非常强壮（如腱），而富含弹性蛋白组织则弹性很好（如血管）。ECM蛋白通过称为**整联蛋白**（integrin）的跨膜受体与细胞作用。它们通过诸如踝蛋白和辅肌动蛋白的桥联蛋白与细胞外侧的ECM蛋白和细胞内侧的肌动蛋白微丝结合。这种相互作用使细胞通过收缩机动蛋白纤维向着ECM的固定结构运动。也有证据表明整联蛋白可激活内部的信号通路，这说明ECM可影响细胞内的基因表达，从而对特殊的ECM组分作出反应从而改变细胞行为。ECM在细胞行为方面功能的一个重要例子是通过基片维持**细胞极性**（cell polarity）（更多例子见Dustin，2002）。在肠上皮中，细胞单层一面上基片的存在负责建立和维持与结缔组织相邻的扁平的基面以及以刷状缘为特征的顶面。

ECM的蛋白聚糖和糖胺聚糖成分行使许多功能。首先，由于这些分子的水溶性极好，可形成水溶胶充任垫子保护组织免受压迫。ECM的蛋白聚糖含量特别高的地方，组织对压力有着很高的耐受力（如软骨）。蛋白聚糖可形成复合体超结构，在其中单个蛋白聚糖分子围绕透明质酸骨架排列。这些复合体可通过储存诸如生长因子的活性分子充当生物储备。事实上，蛋白聚糖对某些信号分子的扩散可能是必需的。例如，Hedgehog和Wingless信号通路在果蝇无糖（sugarless）突变株中都是被抑制的，此突变株不能正确合成蛋白聚糖（Selleck，2000）。最后，蛋白聚糖也有助于介导细胞-基质黏着，这通过与细胞表面称为**糖基转移酶**（glycosyltransferase）的酶结合完成，糖基转移酶的功能是向糖类添加糖基。缺乏游离糖时，糖链被此酶占据。有糖供给时，反应完成，糖链被释放。这样的黏附和释放循环可能对细胞迁移控制特别重要。

3.3 发育概述

发育一词应用于动物时是指一个单细胞成长为成熟个体的过程。通常为了方便将动物发育分为**胚胎**（embryonic）期（在此期间所有重要的器官系统形成）和主要由生长和精细化组成的**胚胎后期**（post-embryonic）（在哺乳动物时）。发育生物学家倾向将精力集中在胚胎期，因为这是最令人激动和剧烈变化的事件发生的时期，但这并不减少胚胎后期发育的重要性。一旦其基本机体框架被建立，何时发育停止还不清楚。在人类，有一个连续的胚后生长期和巩固期，从出生开始持续到最多达其后的二十年，一些器官比别的器官更早达到成熟。当个体达到性成熟时发育是否停止还值得讨论，但许多组织终生需要补充（例如皮肤、血液、肠上皮），这种情况下发育从未真正停止，只是达到一个平衡。许多科学家甚至把衰老视为发育的一部分，因为它是生命周期的一个自然部分。

发育是一个渐进的过程。受精卵最初发育为具有相对粗糙特征的原始胚胎。随着发育的进行，细胞类型的数量增长，而且这些细胞的组成也更加复杂。复杂性是逐渐达到的。在分子水平上，发育整合多个影响细胞行为的不同过程。这些过程是相互关联的，可在胚胎的不同部分单独发生或者结合起来。

- ▶ **细胞增殖**（cell proliferation）：通过反复的细胞分裂，使细胞数目增多。在成熟个体中，通过细胞丢失与之平衡。
- ▶ **生长**（growth）：通过大分子的合成，使整个体积和生物量增加。
- ▶ **分化**（differentiation）：细胞结构和功能特化的过程。
- ▶ **模式形成**（pattern formation）：细胞进行组织的过程，首先形成生物体基本的身体框架，然后是不同组织和器官的详细结构。
- ▶ **形态发生**（morphogenesis）：或整个形态的改变。几种基本机制可能与形态发生相关，包括不同细胞的增殖、选择性细胞-细胞黏着或细胞-基质黏着、细胞形状和大小的改变，程序性细胞死亡的选择性应用以及控制细胞分裂对称和分裂水平。

上述所有过程均由基因控制，它指定胚胎中特定的蛋白质在何处、何时合成，以及为此细胞如何进行功能活动。尽管在一个多细胞生物体中细胞的 DNA 序列有较小的差异，但大多数细胞至少都含有相同的基因。因此，为了使发育中胚胎的细胞多样化，必须有差异基因表达。基因表达由转录因子控制，因此发育最终要依赖每个细胞中活化哪些转录因子（例如，Kuo *et al*.，1992）。如节 3.2.2 所述，转录因子的活性可通过细胞间的信号进行调节。阐明控制生物体发育的信号通路和调控程序是生物医学研究的主要目标。这一章我们要集中讨论脊椎动物和特殊哺乳动物的发育。我们关于人类早期发育的知识不完整，因为研究标本的获取常受伦理和实际因素的限制。因此，许多可获得的信息来自于**发育动物模型**（animal models of development）（框 3.3）。

框 3.3　发育的动物模型

发育的研究集中于相对较少的模式生物上，它们被认为是多细胞生物的主要分类学区分的代表。这些生物中有些最初被选择是因为它们易于获得及圈养，而另一些动物被选择是由于具有特殊的实验优势。对这些生物的研究显示发育基因甚至整个调控路径都是高度保守的。它们已成为与发育有关的人类基因的鉴定和特性分析的有力工具。在已经完成或正在进行的基因组计划中的优越性，已经反映了这些物种作为实验模型的重要性（框 8.8）。

无脊椎动物（invertebrate）

主要的无脊椎动物动物模型是果蝇（黑腹果蝇）和线虫（秀丽新小杆线虫）。这两种生物在遗传上都是可以处理的，因此适合于大规模的突变筛查、遗传分析和遗传操作。两种生物的胚胎（及成虫）是透明的并很适合于外科操作。果蝇是在分子水平被详细研究的第一个动物模型，许多支撑早期胚胎发育的分子机制都是通过这一物种来建立的。果蝇同样为神经发生和眼的发育提供了特别有用的模型。秀丽新小杆线虫因其标准的发育程序而引人注目，其特征是具有几乎不变的细胞谱系和可提供神经系统完整的接线图。谱系突变为研究发育过程中的细胞记忆提供了一个好的方法。线虫生殖孔是已完善建立的器官发生模型。其他的无脊椎动物模型包括软体动物、海鞘（有被膜的）及环节动物。

脊椎动物（vertebrate）

脊椎动物模式生物在解剖学水平和分子水平都可作为代表人类发育的模型。家鸡（*Gallus gallus*）和非洲爪蟾（*Xenopus laevis*）深受喜爱，因为这两个物种都可产生在体外发育且便于操作的健全的胚胎。然而这些物种对于遗传分析却没有什么用处，比如非洲爪蟾，主要是因为繁殖周期长和四倍体基因组使遗传分析十分困难。最近大家对一个相关的物种——热带非洲蛙（*X. tropicalis*）感兴趣，它能产生较小但是周期较短而且是二倍体的胚胎。小鼠（*Mus musculus*）在遗传服从

框 3.3　发育的动物模型（续）

性和遗传操作的适宜性（特别是基因打靶，可使特殊的基因失活，见 20 章）方面都有优势，而且其与人类十分接近。但是，由于其胚胎发育在体内，所以外科操作很难实施。斑马鱼（*Danio reiro*）结合了遗传服从性和可操作性，所以代表了也许是最通用的脊椎动物模式生物。

所有的脊椎动物胚胎都经过相似的发育阶段，包括大的卵细胞的卵裂、原肠胚形成、神经胚形成、体节形成和肢芽的形成。它们达到**系统发生形态**（phylotypic）的阶段，此时所有脊椎动物的躯体计划都相同。尽管有这些相似性，但在五种脊椎动物种类间还存在重要的差别，主要是在卵裂期和原肠胚形成期阶段，反映其选择的营养策略，这一点在节 3.7.2 中有更详细的解释。在原胚轴（节 3.5.1）特化方法的使用及其他的发育过程，诸如性别决定（框 3.9）方面，也存在差异。

3.4　发育过程中细胞的特化

3.4.1　细胞特化涉及一系列不可逆的等级决定

发育可以比作是从山边流下的溪流，顺水而下的树叶代表每个细胞（图 3.6）。水流在到达山底之前可能分支很多次，但一片树叶只能跟随一条支流。到达分叉点时，树叶必须选择其中的一条或另一条支流。一旦作了决定，树叶不能再回去选择另一条支流。决定是不可逆的。

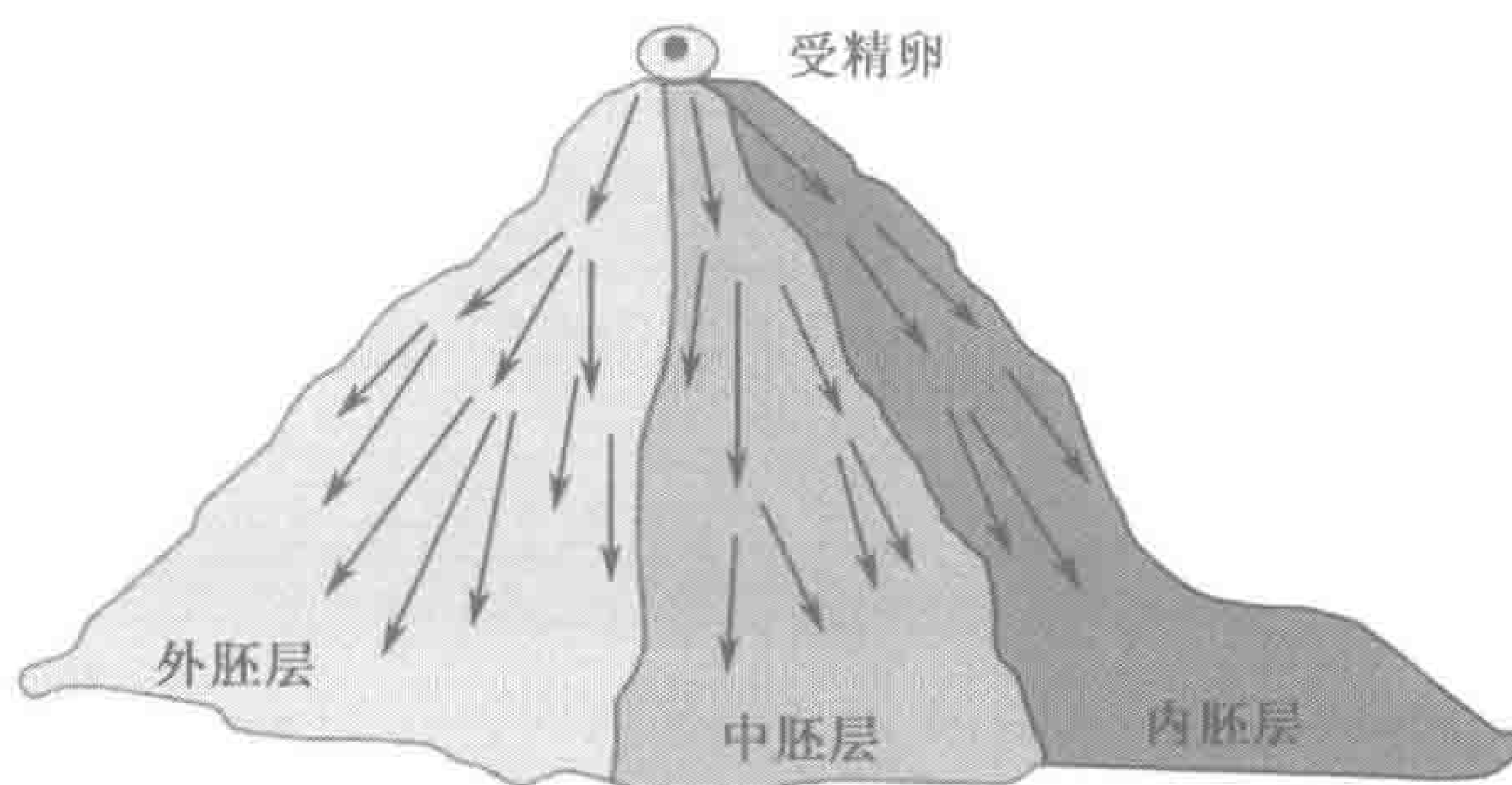


图 3.6　发育途径可以看作山坡上自上而下的溪流分支

受精后，哺乳动物受精卵经历的最初几次细胞分裂是对称的，产生具有相同的发育潜能（或**潜力**，potency）的子细胞。合子和它最近的几个子代细胞未特化，被称为**全能的**（totipotent），因为每一个细胞都保留分化为机体所有可能的细胞的能力，包括胚胎外膜。这类似一片叶子在其山岭之旅的起点。随着发育的进程，细胞变得更特化，同时，其产生不同子代细胞的能力也越来越受限制。细胞被迫做出决定并选择自己的路径。

哺乳动物胚胎的第一个决定是在内细胞团和滋养层之间做出选择。前者会成为合适的胚胎和羊膜，而后者则会发育为绒毛膜和胎盘的胚胎部分。内细胞层的细胞是**多能的**（multipotent），也就是它们能成为胚胎内的所有细胞进而形成一个完整动物体，但是他们不再具有成为来自于滋养层的胚外结构的能力。在这个点之前的任何阶段，胚胎细胞

的潜能都可通过胚胎形成双生子的能力来证实（框3.4）。

框3.4　人胚胎的双生

在人上，大约每200次妊娠就会产生一个双胞胎。有两种不同类型：异卵（二卵，dizygotic）双生和同卵（单卵，monozygotic）双生。异卵双生子来自两个卵细胞独立受精，和其他的同胞相比没有更紧密的关联。尽管在同一个子宫内发育，但胚胎彼此分离，都有独立的胚外膜。同卵双生来自同一次受精事件，是在细胞仍处于全能细胞或多能细胞阶段时胚胎分裂产生。发生于桑椹胚阶段或之前的早期分裂导致胚泡分离，形成由独立的胚外膜包被的胚胎。这种情况大约占同卵双生的三分之一。在剩下的三分之二中，双生发生于胚泡期，这涉及内细胞团的分裂。双生的性质反映了分裂发生和分裂如何完成的准确阶段。大多数情况下，这种分裂发生于怀孕第9天之前，此时羊膜正在形成。此时双胞胎分享共用的绒毛膜腔而被各自的羊膜腔包围。在很少的出生比例中，分裂发生于怀孕9天之后，发育中的胚胎被共用的羊膜包绕。经历不完全分离或者随后融合，这样的双胞胎常会形成连体儿。

已分流进入内细胞团谱系的细胞接着面临三种选择，它们可以选择胚胎三种基本的细胞类型之一：**外胚层**（ectoderm）、**内胚层**（mesoderm）或**中胚层**（endoderm）。每一个**胚层**（germ layer）随后都可以产生特定的、有限范围的细胞类型，但是都不能发育为完整胚胎，它们被称为**多能的**（multipotent）。另外，一旦细胞成为某一胚层，它们通常不能产生具备其他胚层特征的细胞类型，尽管最近有证据显示干细胞可能表现出比曾经预想的更多的发育可塑性（节3.4.6）。例如，外胚层谱系的细胞可以成长为表皮、神经组织和神经嵴细胞，但它一般不能成为肾细胞（来自于中胚层谱系）和肝细胞（来自于内胚层谱系）（框3.5）。最终，在（即三个胚层），祖细胞生成的只是单一型分化的细胞。它们被称为**单能的**（unipotent）。

组织学教科书列出了人体中超过200种不同类型的细胞，包括广泛的功能类型（框3.6）。一些细胞只具有一个器官系统的功能（如肝细胞、心肌细胞），另一些细胞可能具有更普遍的功能，如成纤维细胞。一些成熟、特化的细胞不分裂，称为**终末分化**（terminally differentiate）的细胞。其他的细胞分裂活跃，作为终末分化细胞的前体。这些细胞通常用-blast后缀加以区别，如成骨细胞、成软骨细胞及成肌细胞等。在某些情况下，前体细胞也能经历自我更新，称之为**干细胞**（stem cell）（节3.4.3）。

3.4.2　命运的选择可能依赖于谱系或位置

发育生物学家经常会提出一个问题，一个细胞的命运（可产生的细胞类型范围）是否依赖它的**谱系**（lineage）（即细胞的来源）和它的**位置**（position）（即细胞与谁相连接）。在脊椎动物胚胎的早期发育过程中，细胞的位置似是细胞命运的最重要的决定因素。细胞的命运通常由来自于邻近细胞的信号所决定，称为**诱导**（induction）过程。非洲爪蟾胚胎神经板的形成提供了一个很好的典型的例子（Harland，2000；Bainter *et al.*,2001；图3.7）。

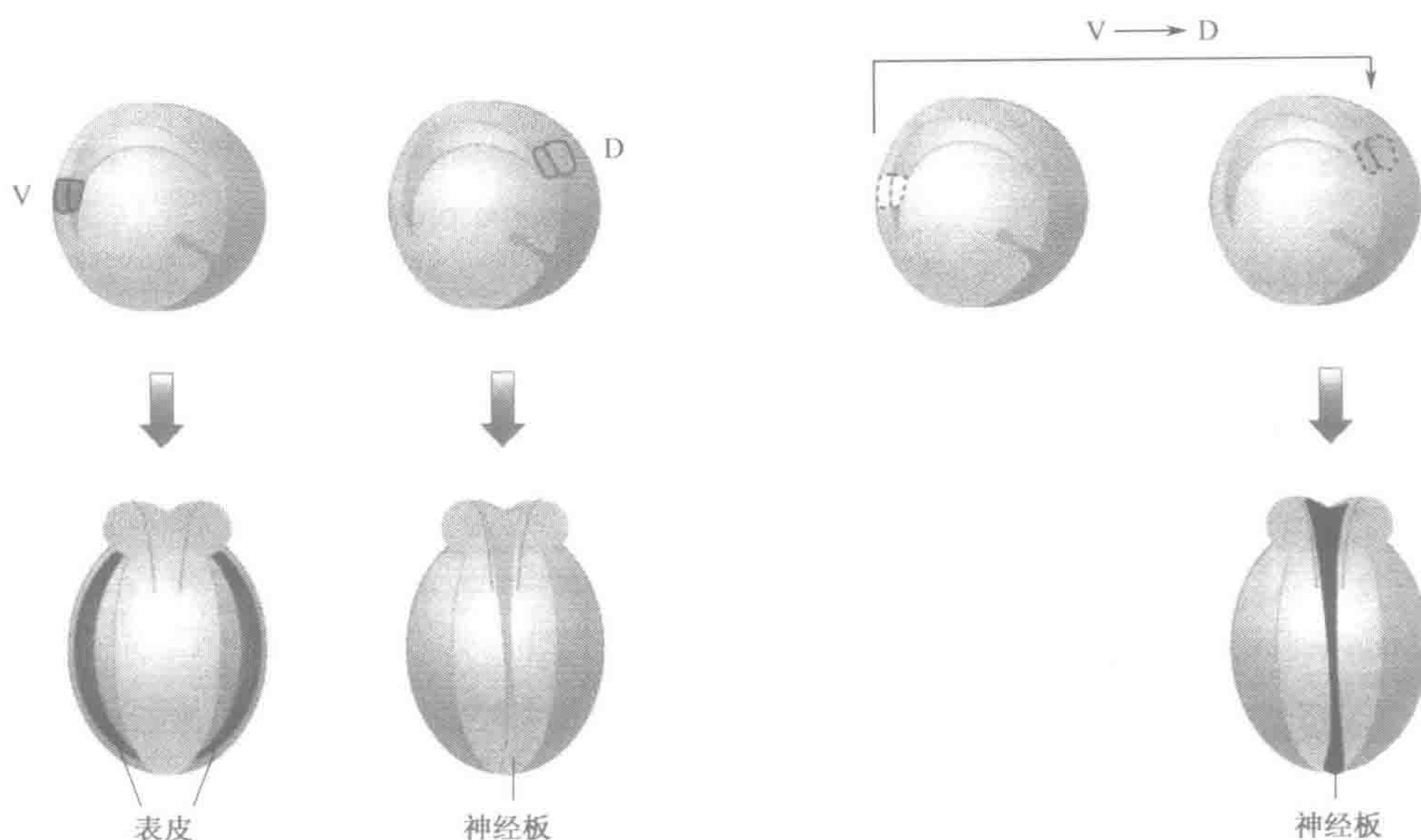


图 3.7　非洲爪蟾的移植实验显示外胚层细胞何时确定成为表皮和神经细胞的命运

左图中，腹部（V）外胚层生成表皮，而背部（D）外胚层生成神经板。在右图中，如果腹部外胚层被移植到胚胎的背部，则可再次特化生成神经板。这表明外胚层细胞的发育命运主要取决于临近的神经板细胞（如轴中胚层细胞）的信号。

神经板来自沿胚胎背中线的表层外胚层，而背中线两侧成为表皮。然而，刚开始表层外胚层是未定型的或原始的—即它有能力成为表皮或神经板。神经诱导的信号来自于**脊索**（notochord），脊索是沿着胚胎前后轴分布的中胚层结构。脊索上的外胚层细胞能接到促进神经发育的信号，但周围的外胚层却不能。然而，如果脊索被切除，那么所有的外胚层都会形成表皮。同样，如果移植一块脊索到腹面或侧面外胚层（正常应该形成表皮），则会形成神经板组织代替。从胚胎的预定的神经板区域移植一块外胚层到任何部位，都会生成表皮，而从胚胎腹侧移植外胚层到脊索上部，将会生成神经板。因而很显然，外胚层的命运——表皮的或神经的——是依赖于其位置而不是其谱系，且最初是可逆的。在这一阶段，细胞的命运是特定的，意味着仍可通过改变细胞所处的环境来改变。经过特定的一段时间，外胚层的命运变得稳定起来，不能通过移植来改变。在这一阶段，细胞认为是被决定的，是不可逆地交托给它们的命运。决定通常意味着细胞开始了一个不可逆的通向分化的分子进程。一些情况下是因为合成了一个新的转录因子，该转录因子不能被激活。另一些情况可能是因为染色质的修饰，在某些部位锁定了基因表达模式，也可能失去了诱导能力。例如，本应发育为表皮的外胚层细胞，可能停止合成对来自脊索的信号起反应的受体。

在脊椎动物胚胎中，细胞命运由谱系决定的例子较少，但一个明显的例子是干细胞行为，我们将在下面讨论。

框 3.5　我们的组织来自哪里——哺乳动物中的发育层次

人类细胞的每一类型通过这一发育层次都可追溯到受精卵。下图显示依照节 3.4.1 类推的下山之旅，从一个受精卵依次形成各种细胞类型所需作出的发育决定。几乎所有的细胞都来自于三个基本胚层之一，主要的例外是原始生殖细胞，它在原肠胚形成之前分离出体细胞系。

合子
卵裂
配子
原始生殖细胞
原肠胚形成
内胚层
膀胱*
尿囊*
肝
胰腺*
气管*
支气管*
肺
原肠
消化管*
甲状腺
咽
咽囊*
I
II
III
IV
中耳*
咽鼓管
扁桃体
隐窝*
原始甲状腺*
甲状旁腺*
甲状旁腺*
腮后小体
外胚层
羊膜和绒毛膜
的胚外外胚层
汗腺*
指甲
乳腺*
头发
躯体外部上皮
皮脂腺*
眼晶状体
听觉小泡*
内耳机制
消化道上皮
口腔上皮
听觉和嗅觉上皮
和嗅觉神经
牙釉质
垂体前叶
肛道上皮
肛管*
运动脊神经根
部分垂体神经
脊索
神经管
视网膜*
和视神经
视觉小泡
脑
颅运动神经
神经脊
颅感觉中枢和神经
感觉脊神经根
脊髓背
根中枢
肾上腺
髓质
牙本质
交感神
经中枢
颅骨和
腮软骨
眼外层
头部间充质
头部连接组织
肌肉
中胚层
脊索
轴旁中胚层
中轴骨
生骨节
四肢骨
芽生附肢
生肌节
四肢肌肉
躯干骨骼肌
皮肤连接组织层
皮肤
附睾输精管
后肾憩室输尿管，
肾盂集合管
中肾管
中肾，
输出管
中段中胚层
原肾
后肾，
肾小管*
Müllerian 管
阴道*
卵巢*
子宫*
侧中胚层
羊膜和绒毛膜
的胚外中胚层
卵黄囊和尿囊
的胚外中胚层
胚体壁中胚层
胸膜
心包
腹膜
内脏中胚层
肾上腺皮质
肠系膜
脏层腹膜
心肌外膜
心外膜，心脏
心肌
间质
或性腺
间充质
成血管组织
脏层胸膜
连接组织和内脏
及血管的平滑肌
血细胞
血管内皮
心内膜

*提示仅上皮部分来源

框 3.6　人类细胞的多样性

组织学教材中描述了超过 200 种细胞类型，这里我们说明一下一些常见的细胞的大小和形态。

卵子（ovum）和**精子**（sperm）（特征见图 3.12，框 11.4 示它们如何从原始生殖细胞产生）

淋巴细胞（lymphocyte）（图 3.9）——小的圆形细胞，直径 6～8μm，胞浆很少，每 μl 血液中通常含 5000 个。

红细胞（erythrocyte）（图 3.9）——直径约为 7.2μm 的双凹圆盘状，红细胞无细胞核、线粒体和核糖体，代谢完全依靠糖酵解。生命周期为 120 天，每 μl 血液中通常含 500 万个。

巨核细胞（megakaryoyte）（图 3.3 和 3.9）——直径 35～150μm 的大骨髓细胞，不规则的分叶核含有大约 8～32 倍基因组，由核内有丝分裂形成。巨核细胞破碎形成成千上万的血小板。

血小板（platelet）（图 3.3 和 3.9）——高度结构化细胞质的 3～5μm 碎片，无核。生命周期为 8 天，每 μl 血液中通常含 200 000 个。

巨噬细胞（macrophage）（图 3.9）——形状可变的细胞，靠伪足运动。通过吞噬作用专门吞噬颗粒，包含许多消化颗粒的溶酶体。许多巨噬细胞常在大的外来物体周围融合，形成巨大的组织细胞。

上皮细胞（epithelial cell）——强壮的黏附细胞挤在一起形成上皮，细胞间存在紧密连接（桥粒）。有些上皮细胞专门负责离子转运、吸收或分泌。

成纤维细胞（fibroblast）——结缔组织的非特化细胞，能分化为软骨、骨、脂肪和平滑肌细胞。

肝细胞（hepatooyte）——直径 20～30μm 的多面体细胞，有时为多核。富含线粒体和内质网，含溶酶体，也可含有脂滴。

肌纤维细胞（muscle fiber cell）（图 3.3）——由肌原细胞融合形成的多核细胞，直径 10～100μm，可以有几厘米长。核排列在外周，大部分内部空间被 1～2μm 的肌原纤维占据，其间有许多线粒体。

神经元（neuron）——形态和大小高度可变，胞体 4～150μm，通常有许多树突和一个轴突。一个细胞可与其他神经元有 10 万个以上的连接。支配脚的脊神经细胞的轴突可长达 1m。

黑色素细胞（melanocyte）（图 3.9）——有长的分支突起的上皮细胞，存在于角质细胞间并将色素包（黑色素体）加入其内。通常每平方毫米的皮肤上有 1500 个黑色素细胞（不论肤色）。

角质化细胞（keratinocyte）——成熟的角质化细胞为天平样结构，富含角质素而缺乏细胞核或任何细胞器。

3.4.3　干细胞是自我更新的始祖细胞

即使当一个生物体完全长大，有些细胞，特别是血细胞、皮肤和肠道上皮细胞以及精子细胞仍需要持续制造。这些细胞是由自我更新的始祖细胞或**干细胞**（stem cell）产生。干细胞可通过正常的对称细胞分裂增殖，产生两个相似的子细胞。当需要时，它们也可经历不对称细胞分裂：一个子细胞拥有与亲代干细胞相同类型的特性，而另一个子细胞改变了特性，并且定型为产生分化细胞的谱系。决定的子细胞的命运不受它的位置或其他细胞信号的影响。干细胞谱系的决定是内在的。这种非条件（自发）特化的细胞命运类型是由于在细胞分裂时调控因子的不均匀分配造成的。例如神经干细胞，存在蛋白质Notch-1（集中在顶极）和 Numb（集中于基极）的不均匀分配。如果在上皮细胞

表面水平面分裂会引起决定因子的平均分配；如果在上皮表面的右角分裂，则导致分配是不均匀的，因而发育不同（Kim and Schagat，1996；图 3.8）。

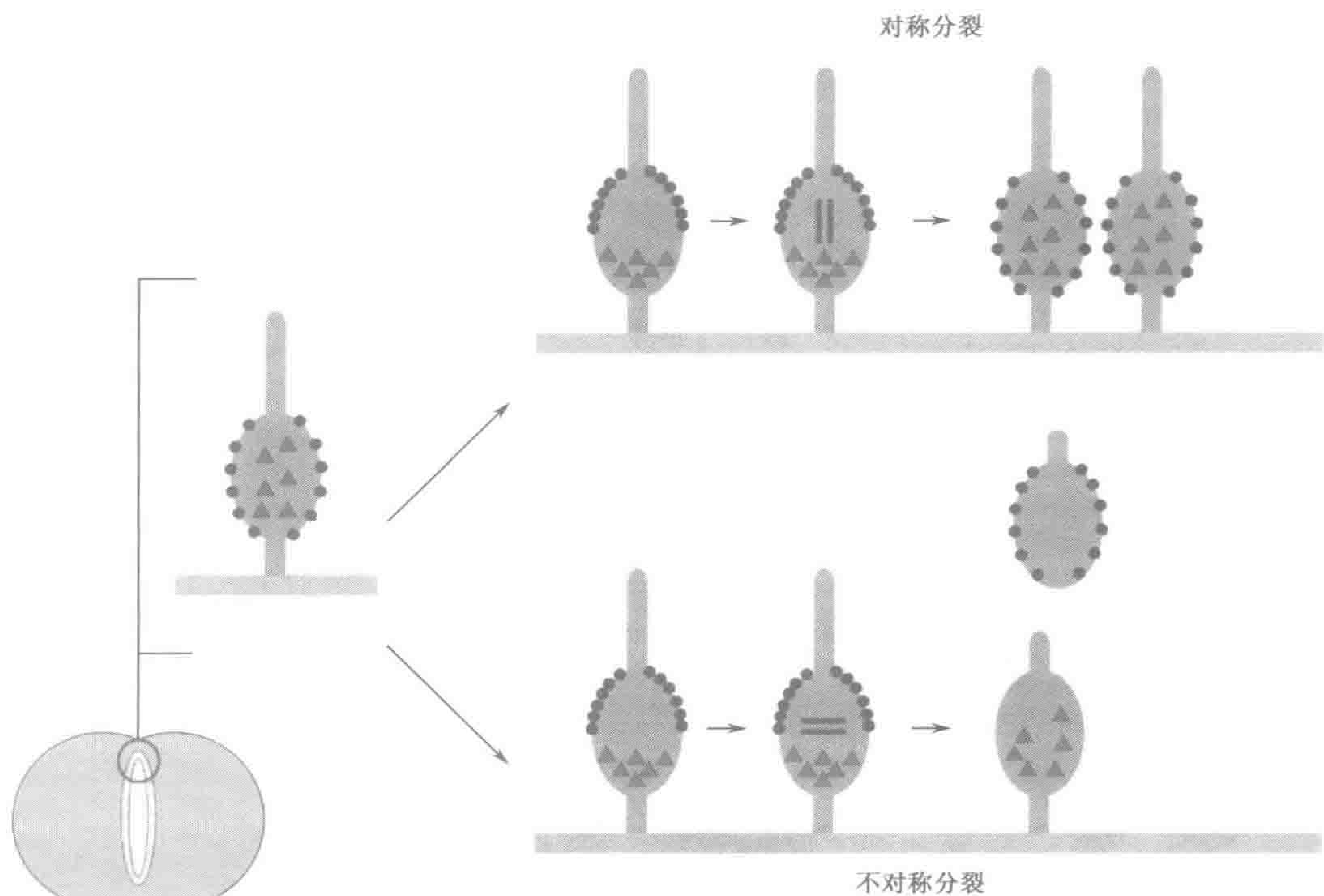


图 3.8　神经干细胞后代的命运取决于细胞分裂面，反映跨膜蛋白（诸如 Notch ）和内细胞决定子（诸如 Numb）的不对称分配

对称分裂发生于神经上皮细胞水平面，导致 Notch（圆圈）和 Numb（三角）在子细胞中平均分配。神经上皮细胞的垂直不对称分裂导致顶部神经元始祖细胞和底部替代干细胞的形成。

血细胞是干细胞谱系的一个有用的例子。一个单独的细胞类型，**造血干细胞**（hematopoetic stem cell，HSC），能生成所有的血细胞（Morrison *et al.*，1995）。这一点可通过一个实验巧妙的证实。小鼠经射线照射，确保其骨髓完全被破坏，然后移植其他品系小鼠的纯化 HSC，输入的细胞能分化和再生血液。HSC 是多能细胞，但它们生成的细胞谱系却逐渐特化，最后产生所有的终末分化血细胞（图 3.9）。

3.4.4　已知存在多种组织干细胞，但关于它们仍有许多东西有待了解

组织干细胞［有时又被称为**躯体干细胞**（somatic stem cell)］是在组织或器官中已分化细胞之间发现的未分化细胞（Verfaile ，2002)。作为干细胞，它能自我更新并分化产生该组织或器官中的主要的特化细胞类型。通常在每一组织中都存在极少量这样的细胞，尽管大多还不知道它们的精确来源，但认为它们定居于每一组织中的特殊区域。有些时候这些细胞多年保持休眠（未分化）状态，直到被疾病或外伤激活，但在另一些情况下，一些常见的基底部需要补充细胞（如替换皮肤和肠的上皮细胞)，它们呈规律性的活化。

第一个成人干细胞是 20 世纪 60 年代发现的，当时发现骨髓至少包含两种类型的干细胞：可生成各种血细胞的造血干细胞和骨髓间充质细胞，骨髓间充质细胞是一种混合

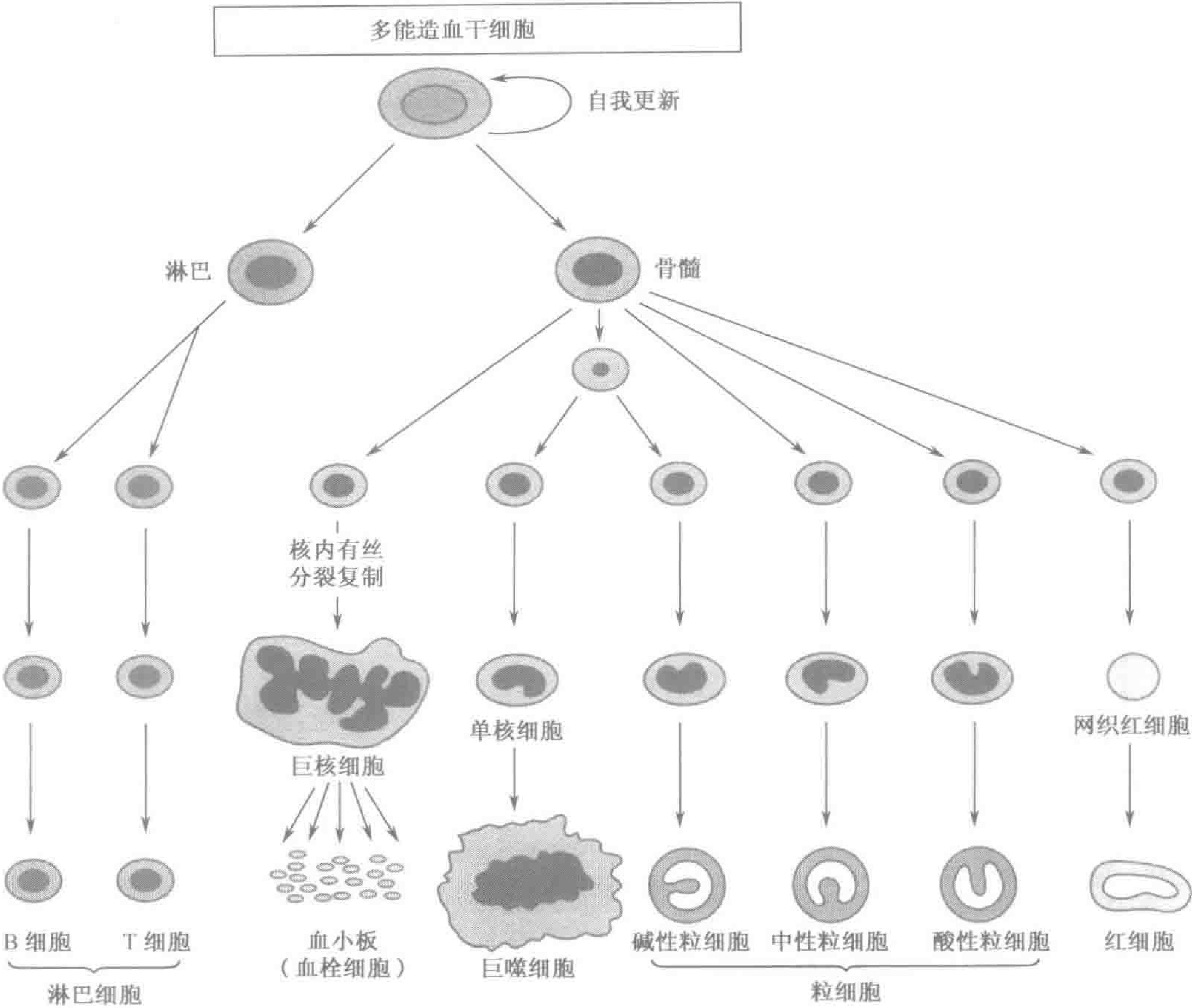


图 3.9　血细胞谱系的决定和分化

细胞群体，可以生成骨、软骨、脂肪和纤维连接组织。此后，又有了一些惊人发现，据报道大脑干细胞能产生所有三种主要的脑细胞类型：神经元（神经细胞）、非神经元星形胶质细胞和少突神经胶质细胞。

已报道多种成人组织含有干细胞（表 3.6），在鉴定一些组织干细胞的特异部位方面也取得了一些进展。使用组织干细胞在实验基础方面的主要困难之一是大部分干细胞在培养中只能保持很短时间的不分化状态（不像胚胎干细胞，见下一节），其分化潜能与胚胎干细胞相比也非常有限。然而最近的证据表明一些成人干细胞的分化潜能比以前认为的可能更具可塑性，这一点将在下节中探讨。

3.4.5　胚胎干（ES）细胞有形成任何组织的潜能

发育极早期的细胞是相对未分化的。顾名思义，**胚胎干细胞**［embryonic stem (ES) cell］来自胚胎。在 20 世纪 80 年代早期取得了重要突破，成功地培养了取自小鼠胚泡内细胞团的细胞（Evans and Kaufman，1981；Martin，1981）。胚胎干细胞是多能的，将其注射入一个不同品系的小鼠的胚泡，再将胚泡移入培养小鼠的输卵管，很快就

能发育为一个健康的成鼠（这一过程的应用将在第 20 章讨论）。

表 3.6　成体干细胞举例

干细胞类型	部　位	分化为
造血干细胞	骨髓	所有血细胞（图 3.9）
骨髓间充质干细胞（间质细胞）	骨髓	骨细胞、软骨细胞、脂肪细胞加上其他类型的结缔组织细胞如腱中的细胞
神经干细胞	脑	神经元（神经细胞）、星形胶质细胞和少突神经胶质细胞
肠上皮干细胞	消化道内层深部隐窝	吸收细胞、杯状细胞、Paneth 细胞、肠内分泌细胞
表皮干细胞	表皮基底层	角质形成细胞
毛囊干细胞	毛囊底部	毛囊和表皮细胞

将内细胞团的细胞转移到含有培养基的培养皿中培养小鼠 ES 细胞，在其内表面覆盖一层**饲养细胞**（feeder cell）。内细胞团（ICM）细胞在饲养细胞的黏性表面可以生长和分裂。饲养细胞顾名思义，能向培养基中释放养分。当内细胞团细胞增殖后，轻轻将其移出，重新接种入几个新的培养皿中进行**亚培养**（subculture）。经过六个月左右的反复亚培养，最初的 30 个左右的 ICM 细胞可产生数百万 ES 细胞。在培养基中增殖六个月甚至更长时间而不分化，仍保持全能性和正常遗传特性的 ES 细胞，被称作**胚胎干（ES）细胞系**。

可采用几种实验方法验证胚胎干细胞是否可能具有全能性，有些方法可产生想得到的分化产物类型。

- **自发分化**（spontaneous differentiation）。正常情况下胚胎干细胞在特定的条件下生长，以保持其处在不分化状态。改变培养基的条件可使细胞互相黏附形成细胞团，称为**胚状体**（embryoid body），之后细胞开始自发分化，但是以一种相当不可预测的方式形成不同类型的细胞。
- **定向分化**（directed differentiation）。细胞被控制（通过改变培养基的化学成分等）使其能分化为特定的预期细胞类型。
- **畸胎瘤形成**（teratoma formation）。将细胞注入免疫抑制小鼠体内，验证称为**畸胎瘤**（teratoma）的特定类型的良性肿瘤的形成。自然发生和实验诱导的畸胎瘤通常包含了许多分化或部分分化细胞类型的混合物——说明 ES 细胞能够分化为多种细胞类型。

注意尽管 ES 细胞也称为干细胞，但它并不能像组织干细胞那样进行不对称分裂，产生一个新的干细胞和一个准备分化的子细胞。胚胎干细胞只能对称分裂，其子细胞具有相同的分化潜能，其潜能取决于实验环境。

最近，已有人类 ES 细胞培养成功的报道（Thomson *et al.*,1998）。ES 细胞来自一家体外受精（*in vitro* fertilization, IVF）诊所的由体外受精的卵子发育而成的胚胎。体外受精方法是为了帮助怀孕有困难的夫妇，通常会有多余的受精卵，可以捐赠用于研究目的。用于获取 ES 细胞的胚胎一般是 5 天左右的胚泡，是中空的，显微镜下为球形，

有100个左右的细胞，其中的大约30个细胞组成内细胞团。人类ES细胞的成功培养是一项重要的、激动人心的突破，因为这项技术提高了新的治疗方法的可能性，且为细胞分化的研究提供了新的途径。发病机制是体细胞损失造成的疾病或严重的损伤，可通过**细胞替代策略**（cell replacement strategy）治疗。因为在理论上胚胎干细胞可被设计分化为特定类型的细胞，所以人们对它的期望非常高（第21章）。

另一个可选择的策略是分离**原始生殖细胞**（primordial germ cell）——生殖嵴细胞，正常能发育为成熟配子（见框11.4插图）。它们能在体外培养生成全能的**胚胎生殖细胞**［embryonic germ（EG）cell］，EG细胞的行为方式与胚胎干细胞很相似。人类EG细胞来自于原始生殖细胞或胚胎和5～10周龄的胎儿，Shamblott等是最先培养EG细胞的（1998）。也可见Donovan和Gearhart（2001）的报道。

3.4.6 组织干细胞的分化潜能是有争议的

近些年来对组织干细胞的分化潜能又有了重新的评价，20世纪90年代后期几篇报道提出某些成体组织干细胞好像是多能的。成体组织干细胞分化为多种相当不同类型细胞的明显能力称为**转分化**（transdifferentiation）或**可塑性**（plasticity）。报道的例子有：

- 造血干细胞可分化为三种主要的脑细胞（神经元、少突神经胶质细胞和星形胶质细胞）、骨骼肌细胞、心肌细胞和肝细胞等（例如Petersen *et al*.,1999）；
- 脑细胞可分化为血细胞和骨骼肌细胞，反之亦然（例如Shih *et al*.,2002；Clarke *et al*.,2000；Bjornson *et al*.,1999）；
- 骨髓间充质细胞可分化为心肌细胞和骨骼肌细胞（Orlic *et al*.,2001）。

欢迎这类报道的热情很大程度上反映了**干细胞治疗**（stem cell therapy）的潜能。理论上任何组织或器官的破坏都可被最接近的干细胞所修复，这些细胞甚至可以事先控制（Weissman，2000；Daley，2002）。例如，神经变性类疾病或骨髓损伤患者的造血干细胞很容易获得，可用于重新设计以获得脑细胞来替代损伤的细胞。然而，关于开发转分化治疗潜力的怀疑也在增长（Medvinsky and Smith，2003），也可见第21章关于使用人类干细胞的一般实践和伦理问题。

3.5 发育中的模式形成

尽管分化赋予细胞特定的结构和功能，但这并不会形成一个生物体，除非细胞以一种有用的方式组织。没有组织，我们可能终止于杂合组织的非晶形小团，只有随机分布的肝细胞、神经元和皮肤细胞，不能形成可识别的组织和器官。然而在个体间有许多细微的区别，同一物种的所有成员都易于改变其基本的躯体计划（Wolpert，1996）。在人类和脊椎动物上总体来说，躯体的发育计划是沿头尾轴呈双侧对称分布，而其他动物的主轴（如背腹轴）是从背部到腹部延伸。在体内，某些器官是不对称分布的，确定为左右轴。对于每一个体，机体的各器官和组织必须与这些轴以相同的方式分布，这种模式在发育的很早期就已经出现。后来，确定的模式出现在特定的器官。每只手的五个手指和每只脚的五个脚趾的形成就是很好的例子。组织内的细胞排列是更细节的模式。在发育的过程中，这些模式是逐渐出现的，从最初粗略的胚胎逐渐变得精细，就像一幅画进入到精雕细刻。本节我

们讨论在发育的胚胎中模式形成是如何开始的和涉及的分子机制。

3.5.1　躯体计划的出现依赖于轴的特性和极化

发育从单个细胞开始，细胞必须被极化以使胚胎生成头和尾、背和腹及左和右。在许多动物中，配子形成时卵的分化包含细胞内不同部位的特定分子沉积，这使细胞有了极性。当受精卵分裂时，这些**决定因子**（determinant）被分配到不同的细胞中，因此胚胎有了极性。在其他一些动物中，卵子的对称被来自环境的外部信号打破。例如鸡，鸡蛋旋转进入输卵管中时，胚胎受地球引力的影响形成了头尾向轴。这两种机制蛙都予以应用：卵细胞预先存在的不对称性由母系基因产物的分配决定，而精子进入位点提供另一种定位协作。哺乳动物的对称破坏机制仍不清楚，但可能也涉及精子进入位点（框 3.7）。

框 3.7　哺乳动物胚胎的极化-信号和基因产物

背腹轴（dorsoventral axis）

哺乳动物胚胎不对称的第一个公开信号是内细胞团与一侧胚泡面的分离（Lu *et al.*，2001；Zernicka-Goetz，2002）。内细胞团代表着胚胎内侧极，因此胚内-胚外轴可以确立。内细胞团的胚胎面暴露于滋养层并与之相接，而胚外侧面开口于囊胚腔。这种环境的不同使内细胞团分化为两个细胞层-胚内侧的原始外胚层和胚外侧的原始内胚层。这依次确定胚胎的背腹轴。还不清楚内细胞层是如何在囊胚第一地点变成位置不对称的，但很有趣的是我们注意到被荧光小珠确定的精子进入的位点总是处于胚内胚外边界的滋养层细胞中。

前后轴（craniocaudal axis）

在哺乳动物胚胎前后轴的确定过程中，精子进入的位置可能也起着重要的作用，但仍不清楚前后轴是如何极化的（Beddington and Robertson，1999；Lu *et*.，2001）。受精诱导第二次有丝分裂，第二极体通常是在精子进入位点后被反向拉出的。这就决定了合子的动-植物轴，极体在动物极中。随后的卵裂发生在动-植物轴之间形成胚泡，表明胚泡与合子之前的动-植物轴双侧对称排列。此胚泡的双侧对称轴预示着原条的排列，但不是它的方向。关于哪一端形成头哪一端形成尾取决于一个称为前部内脏内胚层（AVE）的胚外组织区。小鼠的 AVE 最初位于卵柱的顶部，就在原肠胚形成和胚胎节在外胚层的对侧极建立之前，向前后轴的将来的头极旋转。

左右极（left-right axis）

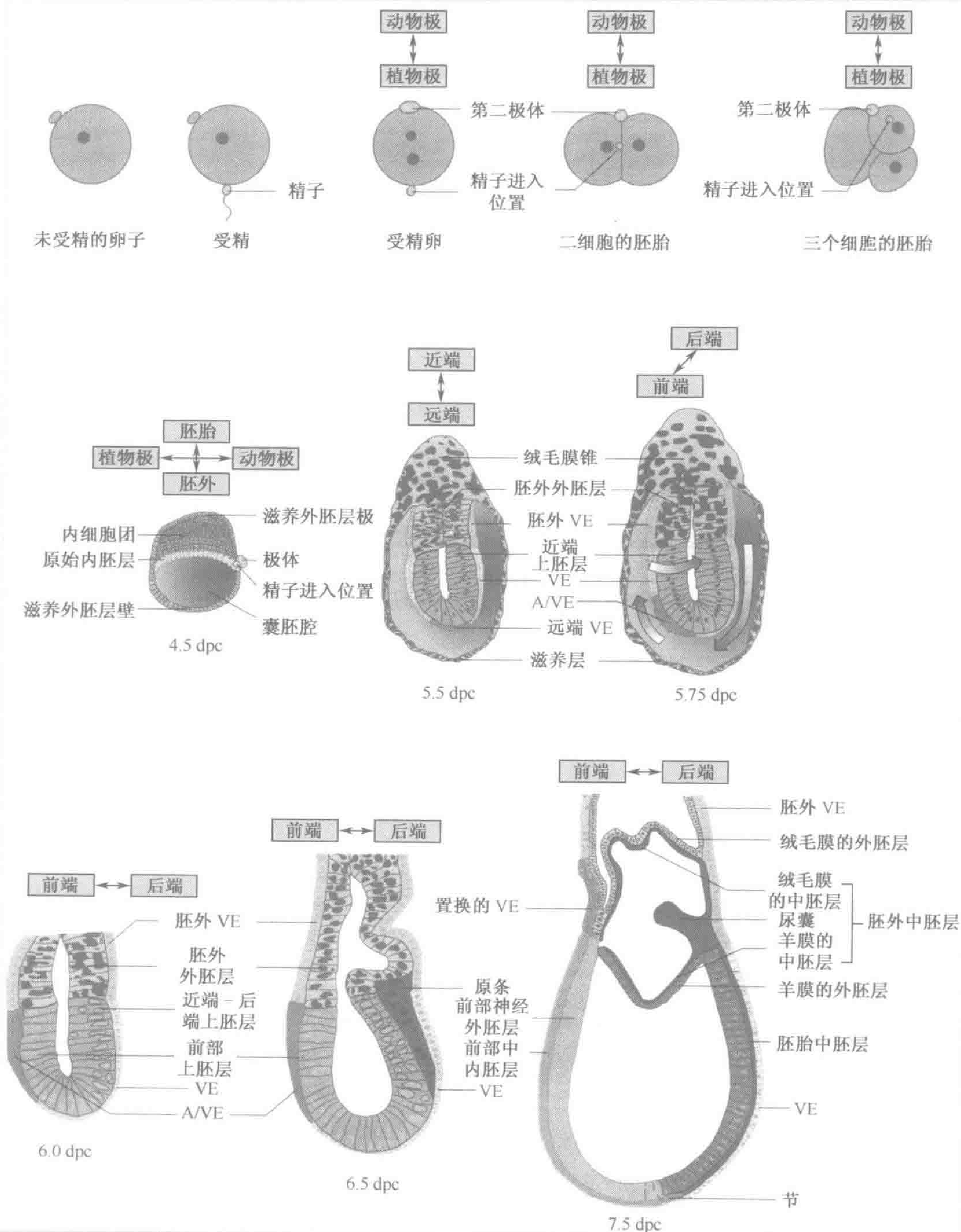
哺乳动物胚胎左右不对称信号首先出现于心管的环化。然而分子不对称要比这出现得早（Capdevila *et*.，2000）。哺乳动物的胚胎主要的决定步骤发生于原肠胚形成期间，这一时期胚胎节中的纤毛旋转会形成特定左右轴需要的单向结周液体流。尽管这一机制还不明确（Tabin and Vogan，2003），但结果是编码胚胎左手侧的信号分子 Nodal 和 Lefty-2 基因的激活。这些初始的信号通路激活左手特定转录因子（例如 Pitx2）而抑制右手特定转录因子。在胚胎的右手侧，Nodal 和 Lefty-2 缺如，Pitx2 未激活。像 Lefty1 等的其他蛋白在胚胎的正中线表达，建立一个能阻止信号从胚胎的一侧向另一侧泄露的屏障。人类基因 *LEFTA*、*LEFTB* 和 *NODAL* 的突变与一系列轴向畸形有关，包括左右逆转（位点逆转，位点不确定）和镜面反射对称（同分异构现象）。这些畸形有时候影响整个机体，而有时候只影响个别器官（内脏异位）。纤毛的单向旋转需要动力蛋白，而影响动力蛋白亚单位的突变也与偏侧缺陷相关。有趣的是这种现象常伴随呼吸道感染和不育症出现，反映了身体其他部位的纤毛不能游动（框 3.1）。

小鼠胚胎中从受精到中条期早期轴形成的概述（封面）

小鼠胚胎中动-植物轴的动物极确定为刚受精后第二极体被挤出的那一点。胚泡中的胚内-胚外

框 3.7　哺乳动物胚胎的极化-信号和基因产物（续）

轴与动-植物轴成直角，由 ICM 的位置决定，胚内极位于胚泡含有 ICM 的一侧，而胚外极位于囊胚腔一侧。卵柱期胚胎的 P-D 轴由位于外胎盘锥的近极和杯形胚底部的远极形成。在原肠胚形成期前，P-D 轴旋转 90°并转化为 A-P 轴。出自 Lu 等（2001）Curr. Opin. Genet. Dev. 11，384～392. 经 Elsevier 许可。



3.5.2 同源异形突变揭示位置确定的分子机制

20世纪80年代进行的大规模果蝇突变形成筛查，发现有一小部分果蝇其机体一部分的发育与另外的某部分很相似［**同源异形转化**（homeotic transformation）］。例如，*Antennapedia* 突变株，会在其头上长出腿而不是触角。这个现象所揭示的关于模式形成的机制就是细胞的位置类别，这个信息告诉每个细胞在胚胎的哪个部位，因而为了形成局部合适的结构应如何行为，这一信息是由基因控制的。这些基因称为**同源异形基因**（homeotic gene）。

随后果蝇的同源异形突变分子分析发现了两簇非常相似的基因（图3.10），每簇基因都编码一个转录因子，该转录因子含有称为同源结构域的保守的DNA结合域。值得注意的是该基因沿着果蝇的头尾轴以一种重叠的方式表达，将果蝇的躯体分为若干不连续的区域。每个区域的特定联合基因表达建立了一个密码，该密码使每个细胞沿着轴都有一个特定的位置身份（Morata，1993；Lawrence and Morata，1994）。通过使一个或多个基因突变或特意的过度表达来操纵密码，可能产生特殊的身体部分转化的果蝇。

在哺乳动物中发现非常相似的同源盒基因簇（*Hox* 基因）。人类和小鼠有四个不连锁的 *Hox* 基因簇沿着头尾轴重叠表达，与果蝇的表达方式有惊人的相似（Krumlauf，1994；Lumsden and Krumlauf，1996；Burke，2000；图3.10）。而且，*Hox* 基因活性的普遍机制似乎是保守的，因为通过敲除突变和特意的表达研究，已获得了小鼠脊柱的身体部分转化。例如，*Hoxc8* 基因靶向破坏可使小鼠多生出一对肋骨，是由于第一腰椎转化为第十三胸椎（Le Mouellic *et.*，1992）。

两个 *Hox* 基因簇，*HoxA* 和 *HoxD* 都沿着肢体重叠表达。敲除和过表达突变这些基因会使小鼠的肢体节段进行特殊的重排。例如，*Hoxa11* 和 *Hoxd11* 基因靶向破坏的小鼠会缺少尺骨和桡骨（Davis *et.*，1995）。人类 *HOXD13* 基因自然发生的突变与一组包括多指畸形的手缺陷有关（Manouvrier-Hanu *et et.*，1999）。

3.5.3 模式形成通常依赖信号梯度

轴的特性形成和极化是发育过程中重要的早期事件，这是因为如果细胞要产生合适的躯体计划，胚胎不同部分的细胞必须最终以它们合成的基因产物的方式的不同而有所不同。只有一个细胞能够告知相对其他细胞在什么位置，才能适当地行动。这又需要一个参考框架。胚胎的主轴能提供参考，使任何一个细胞绝对、准确地确定位置。

细胞是如何知道它们沿着轴的位置并采取相应的行动？这个问题很好，因为不同位置但功能相同的细胞经常要形成不同的结构。例如，在手的发育过程中，来自同一个细胞类型的细胞要形成不同的手指，在体节中来自相同细胞类型的细胞要发育为不同的椎骨（有的有肋骨，有的没有）。

在许多发育系统中，细胞的区域的特异行为已被证实是取决于信号梯度，在不同的信号浓度下相同的靶细胞受到不同的影响。以这种方式作用的信号分子称为**形态生成素**（morphogen）。在脊椎动物的胚胎中，就是用这种机制形成躯体的前后轴及肢体的前后和靠近远端的轴（Wolpert，1996；Ng *et.*，1999）。

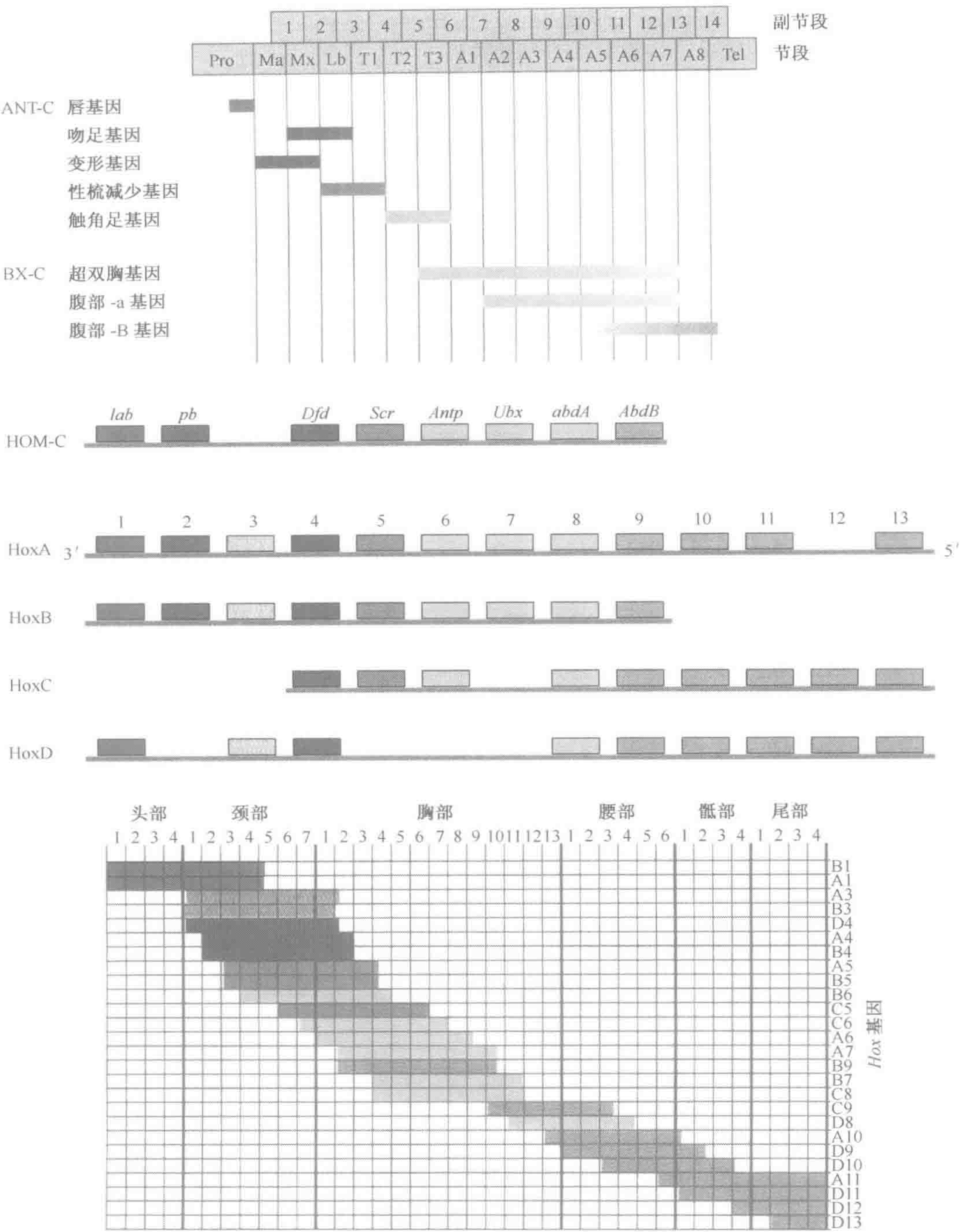


图 3.10　果蝇和小鼠的 HOM-C/*Hox* 基因复合体以及它们沿着胚胎前后轴的表达结构域的比较

出自 Twyman (2001). Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

在发育的肢体中（Schwabe *et*.,1998；Niswander，2003），在每个肢芽的后缘都有一个特定的细胞亚群，称为**极化活动区**（zone of polarizing activity ，ZPA），是形态发

生素梯度的来源（图 3.11）。最靠近 ZPA 的细胞形成手或脚最小最靠后的指（趾），而最远的细胞形成拇指或大脚趾。如将供体 ZPA 移植到已有自己 ZPA 的肢芽前部，可很容易地证实 ZPA 的组织能力。在这个实验中，肢体变得对称，两端都有后部的指（趾）头。发育中的肢体的形态发生素似乎是信号蛋白 **Shh**（sonic hedgehog），尽管这种蛋白被认为是直接发挥作用，因为它不能从它的发源地扩散超过几个细胞的宽度。吸入了 Shh 蛋白的微珠可在功能上替代 ZPA，就像吸入了维甲酸的微珠一样，已知它可以诱导 *Shh* 基因表达。在 ZPA 的中心，*HoxD* 基因远侧的所有的五个基因（*Hoxd9*～*Hoxd13*）全部表达。然而，随着信号强度的逐渐减弱，*HoxD* 基因也被一个接一个的关闭，直到肢芽前部边缘拇指形成的位置，只有 *Hoxd9* 保持打开。通过这种方式，确定主要胚胎轴的信号梯度与控制区域细胞行为的同源基因相联系。

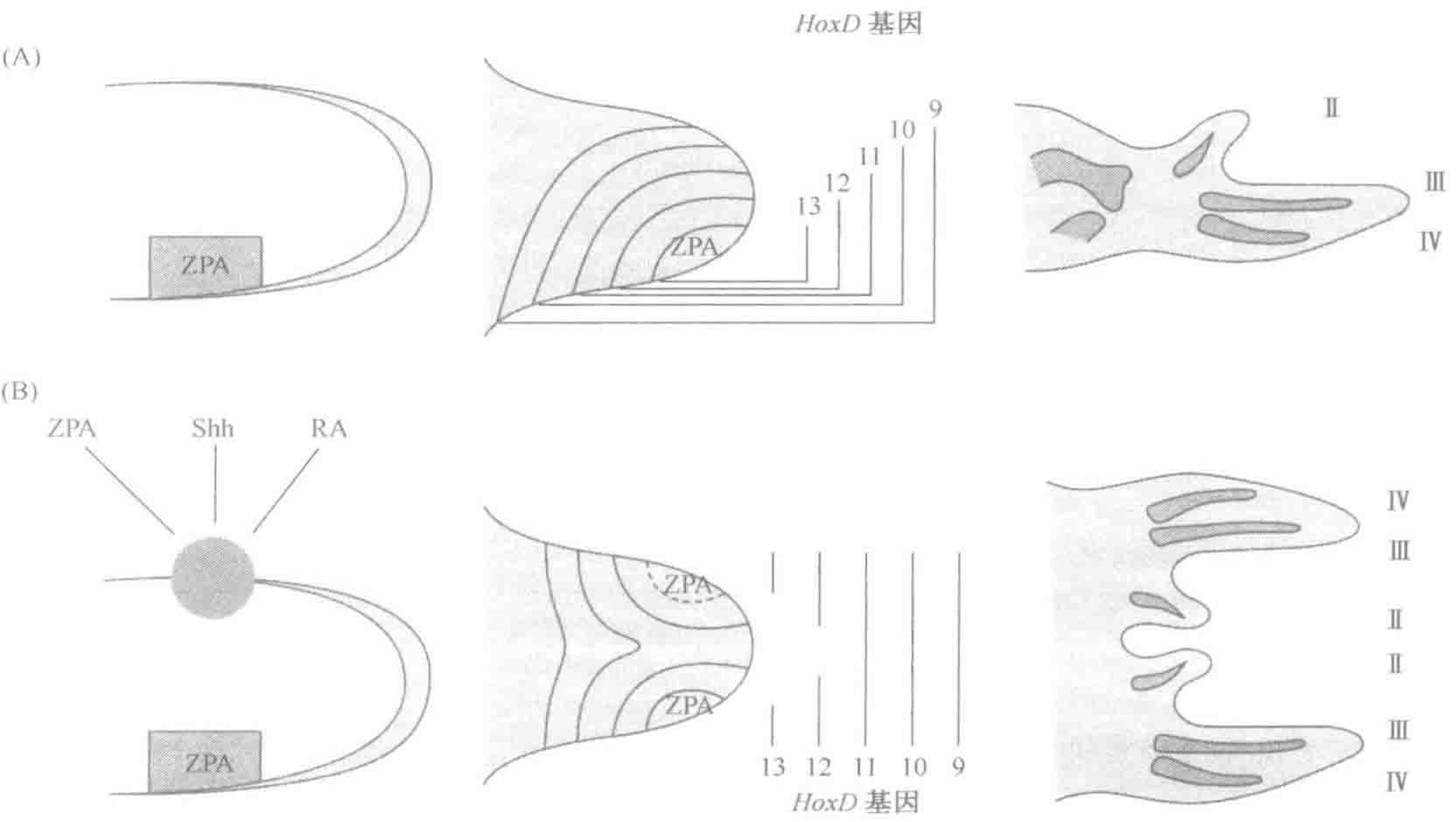


图 3.11　移植第二个极化活动区到小鸡肢芽的前部边缘导致趾的图像复制

（A）"Sonic hedgehog" 在极化活动区（ZPA）表达建立的信号梯度，在肢芽的发育过程中产生了 *HoxD* 基因表达模式的嵌套的重叠模式，导致五趾的特化。（B）ZPA 移植建立了一个相反的梯度，导致了趾的图像颠倒命运。这种效应可被包被了维甲酸（RA）或 "Sonic hedgehog（Shh）" 的微珠所模仿。出自 Twyman（2001）. Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

如上所述，*Hox* 基因不仅在肢体表达，而且沿着胚胎主要的前后轴表达。在这种情况下，引导 *Hox* 基因表达的形态发生素梯度的来源被认为是胚胎节，而形态发生素本身被认为是维甲酸。随着逆行，胚胎节分泌的维甲酸逐渐增加，因此后部的细胞接触的化学物质比前部的多，导致胚胎后部更多的 *Hox* 基因逐步地激活。

3.6　形态发生

随着进行性模式形成和细胞分化，细胞分裂将最终生成一个各类型细胞有组织的胚

胎，但是胚胎并不是一个静态的细胞球。真正的胚胎是动态的结构，细胞和组织不断的相互作用和重排以形成一定的结构和形状。细胞形成层、管、松散的网状团块和致密团块。细胞单独或者成团迁移。一些情况下，这种行为是对发育程序的反应，而另一些情况下，这些过程促进发育，使细胞群聚在一起，否则这些细胞永远不会有联系。几种不同的机制成为形态形成的作用基础，这一点在表 3.7 中总结，下面将进行详细的讨论（Hogan，1999；Mathis and Nicolas，2002；Peifer and McEwan，2002；Lubarsky and Krasnow，2003）。

表 3.7 发育中的形态发生过程及发育系统模型举例

过　程	举　例
细胞增殖分化率	通过进展区细胞增殖，脊椎动物肢芽选择性长出
有丝分裂纺锤体的可选择性定位和/或方向	动物中不同的胚胎分割模式。套用线虫细胞分裂
细胞大小改变	细胞扩展如脂肪细胞贮积脂滴
细胞形态改变	在鸟类和哺乳动物的神经管封闭过程中，从柱形变为楔形细胞
细胞融合	哺乳动物中滋养层细胞和肌管的形成
细胞死亡	脊椎动物肢芽中趾（指）的分离。哺乳动物神经系统中功能性突触形成的选择
细胞-细胞黏附获得	脊椎动物肢芽软骨间质的浓缩
细胞-细胞黏附丢失	哺乳动物原肠胚形成过程中细胞从外胚层分层
细胞-基质相互作用	神经嵴细胞和生殖细胞的迁移；轴突迁移
细胞-基质黏附丢失	细胞从表皮底层分层

出自 Twyman（2001）. Instant Notes in Developmental Biology. © 2001 BIOS Scientific Publishers。

3.6.1 通过改变细胞的形状和大小可驱动形态形成

细胞骨架的重组可引起细胞形状和谐的改变，这对整个的组织结构会产生重要的影响。脊椎动物发育的一个重要里程碑——神经管的形成，部分原因就是由细胞形状的改变所驱动。局部微丝的收缩引起顶部压缩导致神经板的柱状细胞变成楔形，使其发挥枢纽作用。结合神经板边缘的加速增殖，这为整个神经板卷成管状提供了充足的动力（Schoenwolf and Smith，1990）。任何扁平的细胞层中相似的行为将导致细胞层凹陷。

3.6.2 胚胎中主要形态形成的改变来自不同的细胞亲和力

作为组织细胞成为组织或保持组织边界的机制，选择性细胞-细胞黏附和细胞-基质黏附已在节 3.2.3 和节 3.2.4 中介绍。在发育过程中，调整特定细胞黏附分子的合成可使细胞与细胞之间形成连接或断开连接，形成非常动态的重组。在节 3.7.5 中介绍的原肠胚形成或许是最生动的关于形态形成过程的例子。外胚层的单层自己向里弯曲，并转化为胚胎的三个基本胚层，这一过程由细胞形态改变、选择性细胞增殖和细胞亲合力不同三种因素的联合作用而驱动。各种不同过程可能来自细胞黏附特性的改变（McNeil，2000；Irvine and Rauskolb，2001）。

- **迁移**（migration）：胚胎中单个细胞相对于其他细胞的运动。有些细胞，特别是神经嵴细胞和胚芽细胞，在发育的过程中经常广泛的迁移，转移到离它初始位置很远的胚胎部分。
- **移入**（ingression）：细胞从胚胎表面移入内部的运动。
- **移出**（egression）：细胞从胚胎内部移到外面的运动。
- **分层**（delamination）：细胞脱离上皮层的运动，经常由单层细胞转化为多层。在哺乳动物原肠胚形成中这是一个主要的过程。细胞能从基础膜中分离出来，就像皮肤发育中发生的一样。
- **插入**（intercalation）：细胞从多细胞层合并到单层上皮。
- **压缩**（condensation）：松散组织的间质细胞转化为上皮结构。有时称为**间质向上皮的转变**（mesenhymal-to-epithelial transition）。
- **分散**（dispersal）：上皮结构转变为松散的间质细胞——**上皮向间质的转变**（epithelial-to-mesenchymal transition）。
- **外包**（epiboly）：细胞层的伸展。
- **内卷**（involution）：扩展的细胞层向内转动，以使细胞扩展到整个层的内表面形成第二层。

3.6.3　细胞增殖和程序性细胞死亡（凋亡）是重要的形态形成机制

经过一个最初的分裂阶段（胚胎各处所有细胞以大致相同的速度分裂后），其不同部分的细胞开始以不同的速度分裂。这样可以产生新的结构。例如，在侧面中胚层的选择性区域的细胞快速分裂，可以产生胚芽，而其邻近的区域分裂得慢得多，不形成这样的结构。细胞分裂面也是十分重要的。例如，上皮层垂直方向分裂的细胞通过合并新细胞使上皮细胞层扩展。然而，在细胞层同一平面上分裂，会产生另外的一层。就像在节 3.4.3 中讨论的神经干细胞那样，如果细胞是不对称的，细胞分裂平面能影响它产生的子细胞的类型。这种现象出现于女性配子发生过程中，减数分裂时产生了大量的含有细胞质的卵细胞和退化的**极体**（polar body），极体本质上是不需要的单倍体染色体组的废容器。与此相对应，男性配子发生减数分裂时会产生四个完全相同的精子细胞（见框 11.4 中附图）。

程序性细胞死亡（programmed cell death）（**凋亡**，apoptosis）是另一种重要的形态发生机制，因为这能将缝隙引入躯体计划（Vaux and Korsmeyer，1999）。从怀孕大约 45 天开始，手和足板的指（趾）间细胞死亡形成了我们手指和脚趾间的缝隙。在哺乳动物神经系统中，凋亡可剔除那些不活动连接的神经元，使神经元回路逐渐精制形成。惊人的是，高达 50%的神经元通过这种方式处理，在视网膜中可达 80%。

3.7　人类早期发育：受精到原肠胚形成

3.7.1　受精激活卵并使精子与卵子的核融合形成一个独特的个体

受精（fertilization）是两性细胞（**配子**，gamete）融合在一起创造一个新个体的过程，新个体的遗传潜力来自双亲，但与之不同（Wassarman，1999）。男性配子，即精

细胞，是一个细胞质大为减少且有单倍体核的小细胞。由于精子内的正常组蛋白被称为**鱼精蛋白**（protamine）的特殊类型的包装蛋白所取代，因此核被高度浓缩而且转录失活。后端是一根长的**鞭毛**（flagella），可通过摆动提供动力，前端是**顶体囊**（acrosomal vesicle），含有消化酶（图 3.12）。

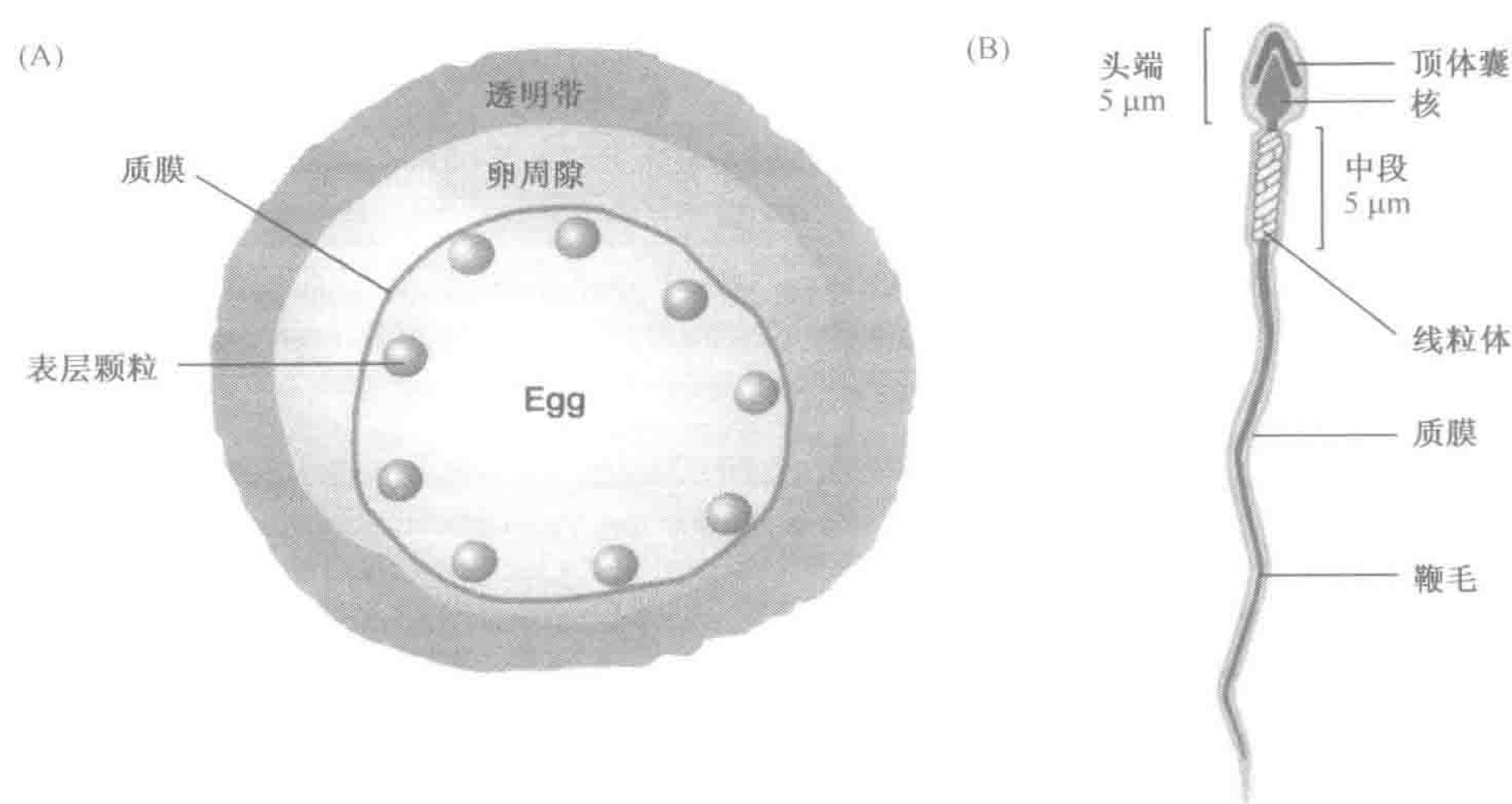


图 3.12　特化的性细胞

(A) **卵子**（ovum）（**卵细胞**，egg）。哺乳动物的卵子是一个大细胞，直径 120μm，被细胞外的包被物透明带所环绕。精子必须首先结合于透明带处。透明带包含三种糖蛋白 ZP-1、ZP-2、ZP-3，这三种蛋白可聚合形成凝胶。减数分裂 I 期的产物第一极体（此处未显示）位于卵周隙内透明带下。在排卵期，卵子处于中期 II，直到受精后，第二阶段减数分裂才完成。受精触发皮质颗粒的分泌，这将有效的抑制其他的精子通过透明带。(B) **精子**（sperm）。细胞远小于卵子，头部为 5μm，包含高度压缩的 DNA；中段为 5μm 的圆柱体，含有许多线粒体；尾部 50μm。在前部，顶体含有帮助精子在卵子透明带上钻孔的酶，允许精子进入并使卵细胞受精。每毫升精液通常含有一亿个精子。出自 Alberts 等（2002）. Molecular Biology of the Cell，4th Edn，p. 1147. Copyright © 2002，Garland Science.

女性配子，即卵细胞（或**卵子**，ovum），是一个含有生长发育初期必需物质的大细胞（图 3.12）。细胞质极度丰富，含大量线粒体和核糖体及大量的 DNA 和 RNA 聚合酶。还有数量可观的蛋白质、RNA、保护性化学物质和形态发生因子。包括鸟类、爬行动物、鱼类、两栖动物和昆虫在内的许多物种，其卵子都包含大量或适度的**卵黄**（yolk）。卵黄集中营养，在胚胎能够独立觅食之前为胚胎的发育提供需要的营养。尽管哺乳动物的胚胎有卵黄囊，但由于其胚胎是靠胎盘血液提供营养，因此哺乳动物卵中不需要卵黄。

质膜的外面是卵黄膜，在哺乳动物中是分离的并且厚的细胞外基质称为**透明带**（zona pellucida）。哺乳动物中，卵也是被称为**卵丘细胞**（cumulus cell）的细胞层所包围，卵丘细胞在排卵前或刚刚排卵后为卵提供营养。

人类的精子必须迁移很远的距离，射精进入阴道的 2.8 亿左右的精子中只有 200 个左右能到达受精发生的输卵管所需要的部位。受精以精子附着于透明带开始，随后精子顶体囊释放酶类，引起透明带的局部消化。然后精子的头部与卵的质膜融合，精子的核

进入卵的细胞质中。在卵细胞内，最初精子和卵的单倍体染色体是各自分离的，分别形成男性和女性的**原核**（pronuclei）。它们随后融合形成二倍体核。受精的卵母细胞称为**合子**（zygote）。

3.7.2 卵裂使合子分为许多更小的细胞

卵裂是合子分裂形成称为**分裂球**（blastomere）的许多小细胞的发育阶段。在不同的物种中，早期卵裂的性质有很大不同，在许多昆虫中（包括果蝇），这个过程甚至不包括细胞分裂［取而代之的是受精卵的细胞核在公共的细胞质中经过一系列的分裂，形成一个大的、扁平的多核细胞，即合胞体胚盘（syncytial blastoderm）］。除了一些例外的情况，卵裂的结果通常是一个细胞球，通常被充满液体的腔所包围称为**囊胚腔**（blastocele）。

在大多数无脊椎动物中，来自于卵裂的细胞球称为**囊胚**（blastula），但在脊椎动物中命名法有所改变。在两栖动物和哺乳动物中，**桑椹胚**（morula）一词指的是来自早期卵裂的原始的、松散集合的细胞球。因此，当充满液体的囊胚腔形成的时候，在两栖动物中这个细胞球被称为囊胚，而在哺乳动物中被称为**胚泡**（blastocyst）（图 3.13）。在鸟类、鱼类和爬行动物中这种情况又有所不同，它们的卵子包含大量卵黄，能阻止细胞分裂。卵裂受到限制，只能在细胞外周形成扁平的胚盘。

许多物种（不包括哺乳动物——见下文）的卵裂期分裂十分迅速，这是因为在细胞周期中，DNA 复制和有丝分裂之间没有插入 G_1 和 G_2 间隙期。在这种情况下，没有胚胎净增长，而只是细胞数目增加，但细胞体积减小。这种情况的出现说明来自合子的基因组（**合子基因组**，zygotic genome）在卵裂期转录失活，因此卵裂期很大程度上依赖于分配入卵子细胞质中的母系基因产物。母系基因产物调控细胞周期并决定卵裂速度，而卵裂分裂是**同步的**（synchronous）。这种调节类型通常称为**母系基因组**（maternal genome）调节，而母系基因产物称为**母系决定子**（maternal determinant）。

有几种情况哺乳动物的卵裂是特殊的：

- **合子基因组的早期激活**——一些物种早在两细胞阶段，结果是卵裂期分裂是由合子基因组调控而不是由母系遗传的基因产物调控，分裂缓慢（因为细胞周期包括 G_1 和 G_2 间隙期）且不同步。
- **旋转卵裂**（rotational cleavage）——第一个卵裂平面是垂直的，但在第二轮细胞分裂中，一个细胞垂直分裂，而另一个是水平的分裂（图 3.13）。
- **压缩**（compaction）——松散连接的八细胞胚胎分裂球相互挤扁以尽量接触，形成一个紧密聚集的桑椹胚（注：压缩确实发生在许多非哺乳动物胚胎中，包括非洲爪蟾，但压缩在哺乳动物中惊人的明显）。压缩有介导一定程度**细胞极性**（cell polarity）的效应。在压缩之前，分裂球是均匀分布微绒毛的圆形细胞团，而且不论细胞之间在哪连接，都可以发现细胞连接分子 E-钙黏着蛋白。压缩之后，情况则十分不同，微绒毛局限于顶部表面，而 E-钙黏着蛋白分布于整个基底外侧表面。现在细胞与它的邻居形成紧密连接，细胞骨架成分重组形成顶带（表 3.4）。

在哺乳动物大约 16 个细胞的桑椹胚阶段，区别两种细胞类型已成为可能：外侧极化细胞和内侧非极性细胞。随着非极性细胞群体的增加，细胞开始通过间隙连接互相交

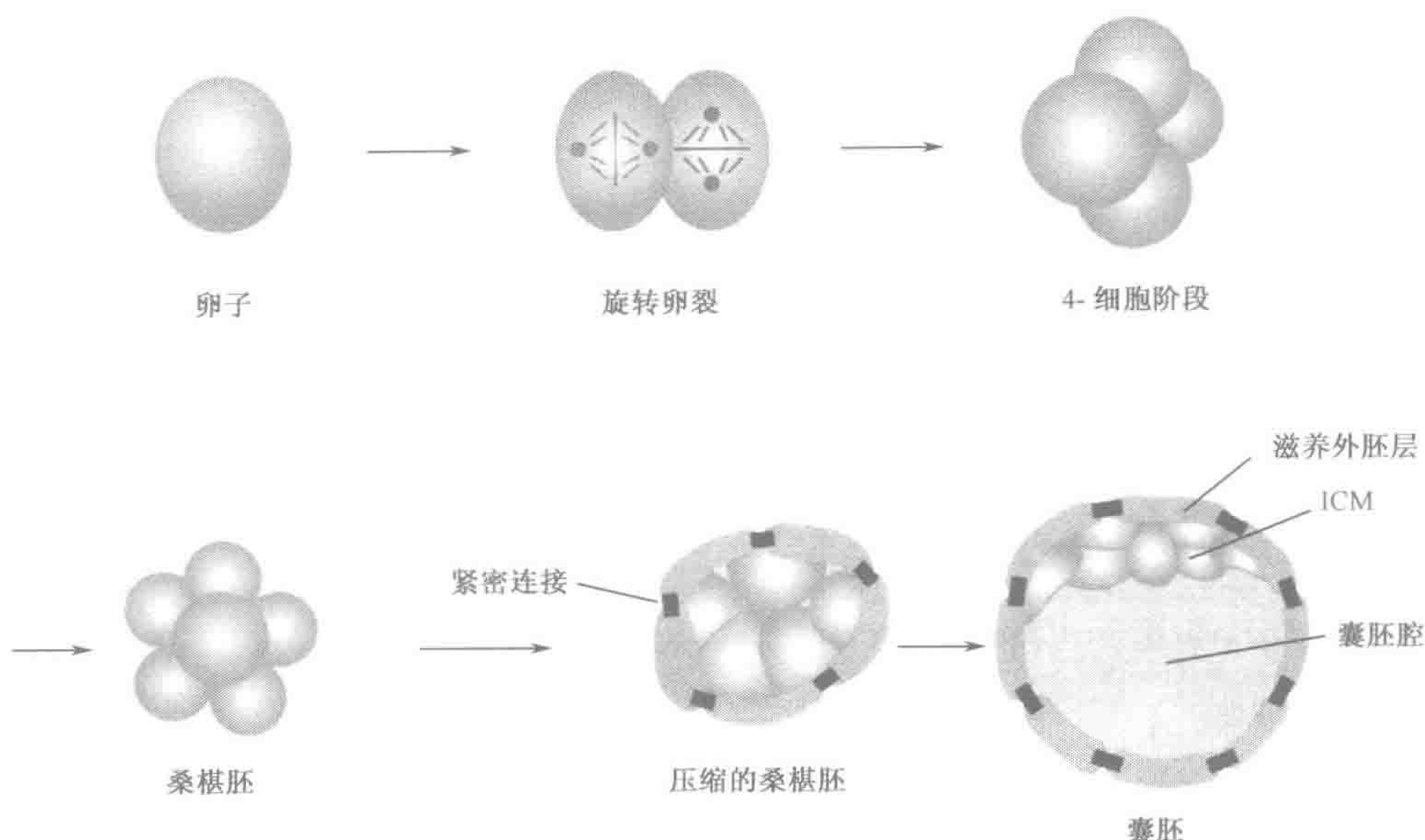


图 3.13　哺乳动物胚胎的早期发育，从受精到胚泡形成

出自 Twyman（2001）. Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

流。这两种细胞间的区别是下一节我们将要阐明的一个基本问题。

3.7.3　哺乳动物早期胚胎中只有很小比例的细胞生成成熟的生物体

在许多动物的发育模型中，生物体由早期胚胎所有细胞的后代形成。然而哺乳动物却十分不同，只有小部分早期胚胎的细胞可以生长为正常的机体。这是因为大部分哺乳动物的早期发育是建立维持生命活动的组织而不是为了形成最终的机体，称为**胚外膜**（extraembryonic membrane）和**胎盘**（placenta）（框 3.8）。

框 3.8　胚胎外膜和胎盘

早期的哺乳动物发育主要集中于组织的形成，其中大部分并不形成最终的生物体*，它们是四层**胚外膜**（extra-embryonic membrane）（卵黄囊、羊膜、绒毛膜和尿囊）和**胎盘**（placenta）——来自于胚胎组织（绒毛膜）和母体组织的联合体。这些生命支持系统要保证胚胎的营养、呼吸和排泄，也起到保护胚胎（和后来的胎儿）的作用。

（* 例外：卵黄囊的背部，作为原肠的前体参与胚胎形成；尿囊在成人体内表现为纤维索，是脐带的残留物。）

卵黄囊（yolk sac）。四个胚外膜中最原始的一个，卵黄囊可见于所有的羊膜动物中，也可见于鲨鱼、硬骨鱼及一些两栖动物。在鸟类的胚胎中，卵黄囊包围着有营养的**卵黄**（yolk mass）（蛋中的黄色部分，大多由磷脂组成）。在许多哺乳动物中（包括人类和小鼠），卵黄囊内不含卵黄。卵黄囊通常很重要是因为：

▶ 起源于上胚层的**原始生殖细胞**（primordial germ cell），在生殖嵴克隆化之前迁移到卵黄囊。

框 3.8 胚胎外膜和胎盘（续）

▶ 卵黄囊是孕体第一批血细胞和第一批血管中的大部分（其中的一些延伸到胚胎中）的发源地。卵黄囊起源于内脏侧板中胚层和内胚层。

羊膜（amnion）。羊膜不仅存在于羊膜类动物中（哺乳动物、鸟类、爬行动物），也以原始形式存在于一些硬骨鱼和两栖动物中。它是四层胚外膜中最里侧的一层，紧邻并包裹着胚胎。它含有羊膜液，使胚胎浸于其中，因此可以防止胚胎在发育过程中过于干燥，使胚胎漂于其中（这样可减少重力对机体的影响），作为一层水垫保护胚胎免受机械震荡。羊膜来自于外胚层和体节侧板中胚层。

绒毛膜（chorion）。像羊膜一样，它不仅存在于哺乳动物中，也以原始形式存在于一些硬骨鱼和两栖动物中。也来自于外胚层和体节侧板中胚层。在鸟类的胚胎，比如鸡，绒毛膜被压至紧贴壳膜，但在哺乳动物胚胎，它是由**滋养层细胞**（trophoblast cell）组成，滋养层细胞能产生一些酶类侵蚀子宫内壁，有助于胚胎植入子宫壁。绒毛膜也是影响子宫和其他系统激素的来源地（绒毛膜促性腺激素）。在所有的情况下，绒毛膜都作为呼吸交换的一个表面。在胎盘哺乳动物中，绒毛膜提供胎盘的胚胎组成部分（下面）。

尿囊（allantois）。胚外膜中最为进化的部分，尿囊只存在于羊膜类动物中（或许应该称之为尿囊类动物）。在鸟类、爬行动物和大部分羊膜类动物中，尿囊是作为一个废物（尿）的储存系统，但在胎盘哺乳动物中却不是这样。尿囊来自于后肠底部的膨出，因此它由内胚层和内脏侧板中胚层组成。尽管在一些哺乳动物中尿囊仍很有作用，但在其他的动物（包括人类）上，尿囊除了它的血管能发育为**脐带**（umbilical cord）之外，只是一个残余物（现在没有任何功能）。

胎盘（placenta）。胎盘只能存在于胎盘哺乳动物中，它部分来自于孕体，部分来自于尿囊。当受精卵植入后，胎盘开始发育，这时胚胎在邻近的子宫内膜诱导一个反应，使之变为一个富含血管组织的营养包，称为**蜕膜**（decidua）。在发育的第二周和第三周，滋养层组织变为空泡状，这些空泡与邻近的母体毛细血管相连接，迅速被血液充满。当绒毛膜形成的时候，它向外生长，绒毛膜绒毛进入空泡，将母体和胚胎的血液紧密联系起来。在第三周末的时候，绒毛膜分化完全，并有可与胚胎连接的血管系统。在绒毛膜绒毛上，营养和废物进行交换。最初，胚胎被蜕膜完全包围，但随着它生长并扩张入子宫的时候，覆盖的蜕膜组织（蜕膜荚膜）脱落，然后分解。成熟的胎盘完全来自于其下的蜕膜基底。

两种细胞类型在哺乳动物 16 个细胞阶段时可以区分开来，它们有不同的命运。外侧极化细胞组成**滋养层**（trophoblast）（或**滋养外胚层**，trophectoderm），它会继续形成为四个胚外膜之一的**绒毛膜**（chorion），为**胎盘**（placenta）提供胚胎的部分。内侧非极性细胞组成**成胚细胞**（embryoblast）。当囊胚腔形成的时候（在人类大约 32 个细胞阶段），内侧非极性细胞聚集在囊胚腔的一端形成偏心**内细胞团**（ICM）。内细胞团的细胞将会生成机体所有的细胞和其他三个外胚膜（框 3.8）。

3.7.4 植入

经过一定的时间（人类发育的第 5 天），胚泡开始**孵育**（hatch）：酶被释放出来，在透明带上钻一个洞，然后胚泡被挤出来。这时胚泡自由的直接与子宫内膜接触。到达子宫后，很快（人类发育第 6 天），胚泡紧紧附着在子宫上皮（**植入**，implantation）。滋养层细胞迅速增殖，分化为**细胞滋养层**（cytotrophoblast）的内层和外侧多核细胞层，即**合胞体滋养层**（syncytiotrophoblast），后者开始侵入子宫的结缔组织中。

甚至在植入发生之前，ICM 细胞就开始分化为不同的外部和内部细胞层。外部细胞层是**上胚层**（epiblast）（**原始外胚层**，primitive ectoderm）。上胚层细胞将会发育为外胚层、内胚层和中胚层（并因此形成所有的胚胎组织），以及羊膜、卵黄囊和尿囊。内细胞层即**下胚层**（hypoblast）（**原始内胚层**，primitive endoderm），将会发育为**胚外中胚层**（extraembryonic mesoderm），排列着原始卵黄囊和囊胚腔。

充满液体的腔——**羊膜腔**（amniotic cavity），在内细胞团内形成，外附羊膜。来自于部分内细胞团的胚胎现在由上胚层和下胚层（被称为**二胚层胚盘**，bilaminar germ disc）组成。它位于两个充满液体的腔之间，一侧是羊膜腔，另一侧是**卵黄囊**（yolk sac）（图 3.14）。

3.7.5　原肠胚形成是一个动态的过程，借此上胚层的细胞生成三个胚层

就像 Lewis Wolpert 的著名论断一样，“在你的生命过程中真正重要的事情不是出生、结婚或死亡，而是原肠胚形成”。**原肠胚**（gastrulation）形成发生于人类发育的第三周，是发育过程中第一个主要的形态发生过程。在原肠胚形成过程中，机体的方向消失，胚胎转化为三胚层结构：**外胚层**（ectoderm）、**内胚层**（endoderm）和**中胚层**（mesoderm）。这三个胚层都来自于上胚层，是生物体所有组织的祖先（框 3.5）。

在哺乳动物、鸟类和爬行动物中，代表原肠胚形成特征的主要结构是一个直线形的物体——**原条**（primitive streak）。这出现于人类发育的第 15 天，如同一个沿着椭圆形二胚层胚盘经度中线的浅沟。在次日的发育过程中，**原沟**（primitive groove）变得更深更长，大约占整个胚胎的一半长度。到第 16 天的时候，一个深的凹陷（**原始小坑**，primitive pit）被上胚层（**原始节**，primitive node）的小丘所包围，十分明显的位于沟的末端，靠近胚盘的中心（图 3.14）。

原肠胚形成的过程是一个极为动态的过程，包含非常快的细胞移动（Narasimha and Leptin，2000；Myers *et*.，2002）。在人类胚胎发育的第 16 天，靠近原条的上胚层细胞开始增殖，变平和失去它们相互之间的连接。这些扁平的细胞长出伪足，使它们能通过原条迁移到上胚层和下胚层之间的空间（图 3.14）。有些进入的上胚层细胞侵入到下胚层并且置换它的细胞，导致下胚层最终被一新的细胞层——**胚内中胚层**（intraembryonic mesoderm）完全取代。从第 16 天开始，许多上胚层细胞通过原条迁移，歧化后进入上胚层和新生的定形内胚层之间的空间形成第三层——**内胚层**（entoderm）。当胚内中胚层和内胚层形成的时候，残留的上胚层就被称为外胚层，新的三层结构称作**三胚层胚盘**（trilaminar germ disc）（图 3.14）。

进入的中胚层细胞按不同的方向迁移，有的向前部，有的向后部，而其他的则沉积在中线上。迁移通过原始小坑和在中线停留的细胞形成两个结构：

- **脊索前板**（prechordal plate）：中胚层前部到原始坑的紧密团块。脊索前盘能诱导像大脑那样的重要的前部中线结构。
- **头突**（notochordal process）：是一个中空的管，源自原始小坑，当原节区的细胞增殖及原条退化时其长度增长，加在它的近端。到人类发育的第 20 天，头突完全形成，但随之从一个中空的管转化为一个固体的杆，即**脊索**（notochord）。脊索将来会诱导神经系统组件的形成（见下节）。

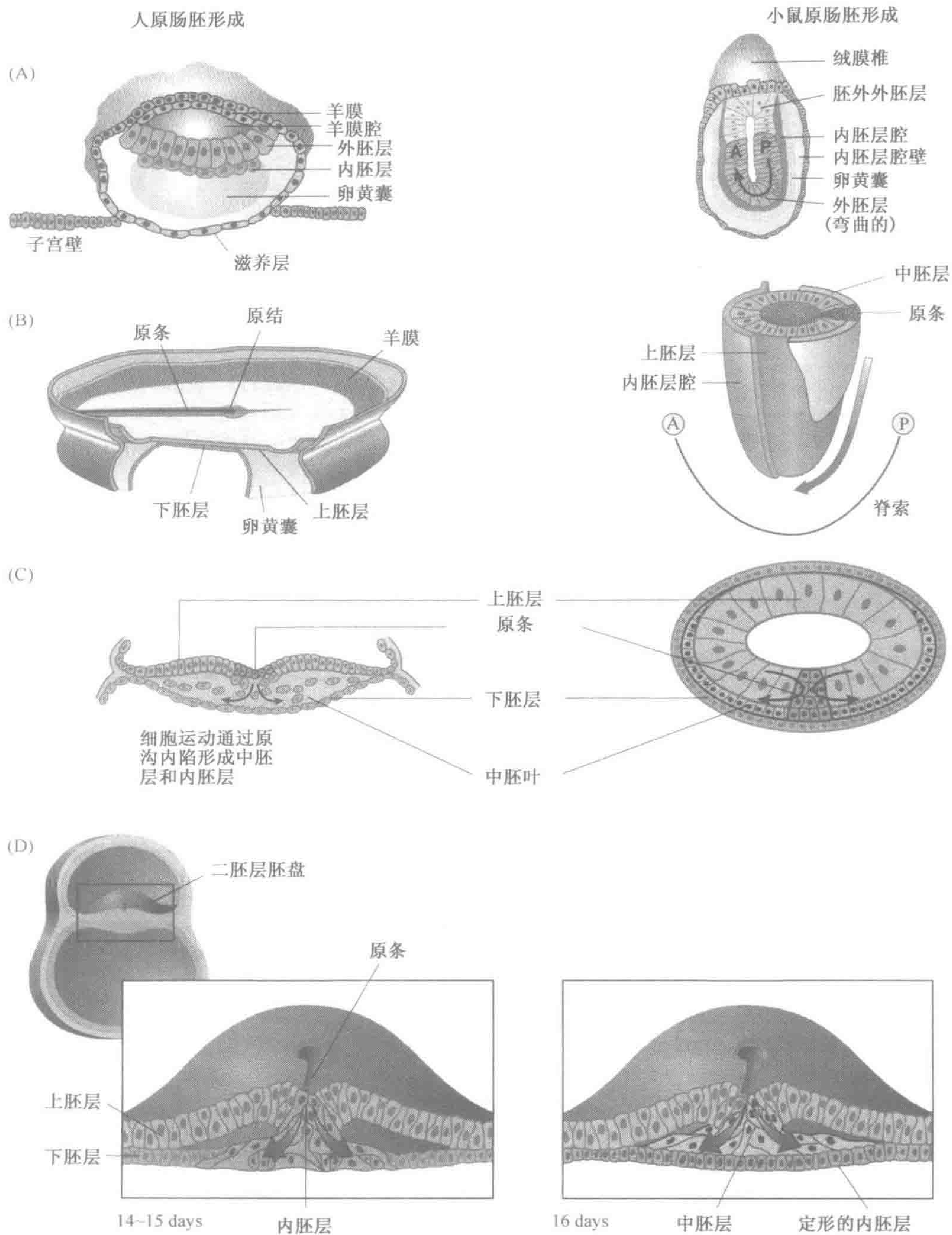


图 3.14　人类和其他灵长类动物的原肠胚形成涉及通过上胚层细胞的程序性分层使一个扁平的二胚层胚盘重组形成一个三层胚胎

(A-C) 由于所有哺乳动物的原肠胚形成的结果都很相似，主要的差别可能在于形态形成的细节上，特别是在胚外结构的形成和应用方式上。右图示小鼠的原肠胚形成作为对照。在这一物种中，扁平的胚盘被杯形的卵柱所取代，细胞从原条移至胚胎的表面而不是移入胚胎内部。将来的前后轴被杯底环绕。(D) 人类发育的 14～15 天之后，进入的中胚层细胞侵入下胚层并替代了它的细胞，导致胚胎内胚层的形成。在第 16 天，一些进入的上胚层细胞分化进入外胚层和新生的内胚层之间的空隙形成胚胎中胚层。**注**：上胚层生成所有的三个胚层（外胚层、中胚层、内胚层），而下胚层生成排列于卵黄囊的胚外中胚层。人胚胎重绘自 Larsen (2002) Human Embryology 3rd Edn，经 Elsevier 许可，鼠胚胎重绘自 Twyman (2001) . Instant Notes in Developmental Biology，由 BIOS Scientific Publishers 出版。

进入的中胚层细胞在脊索的任何一侧可迁移浓缩形成杆状或层状结构。有三种主要结构（图 3.15）。

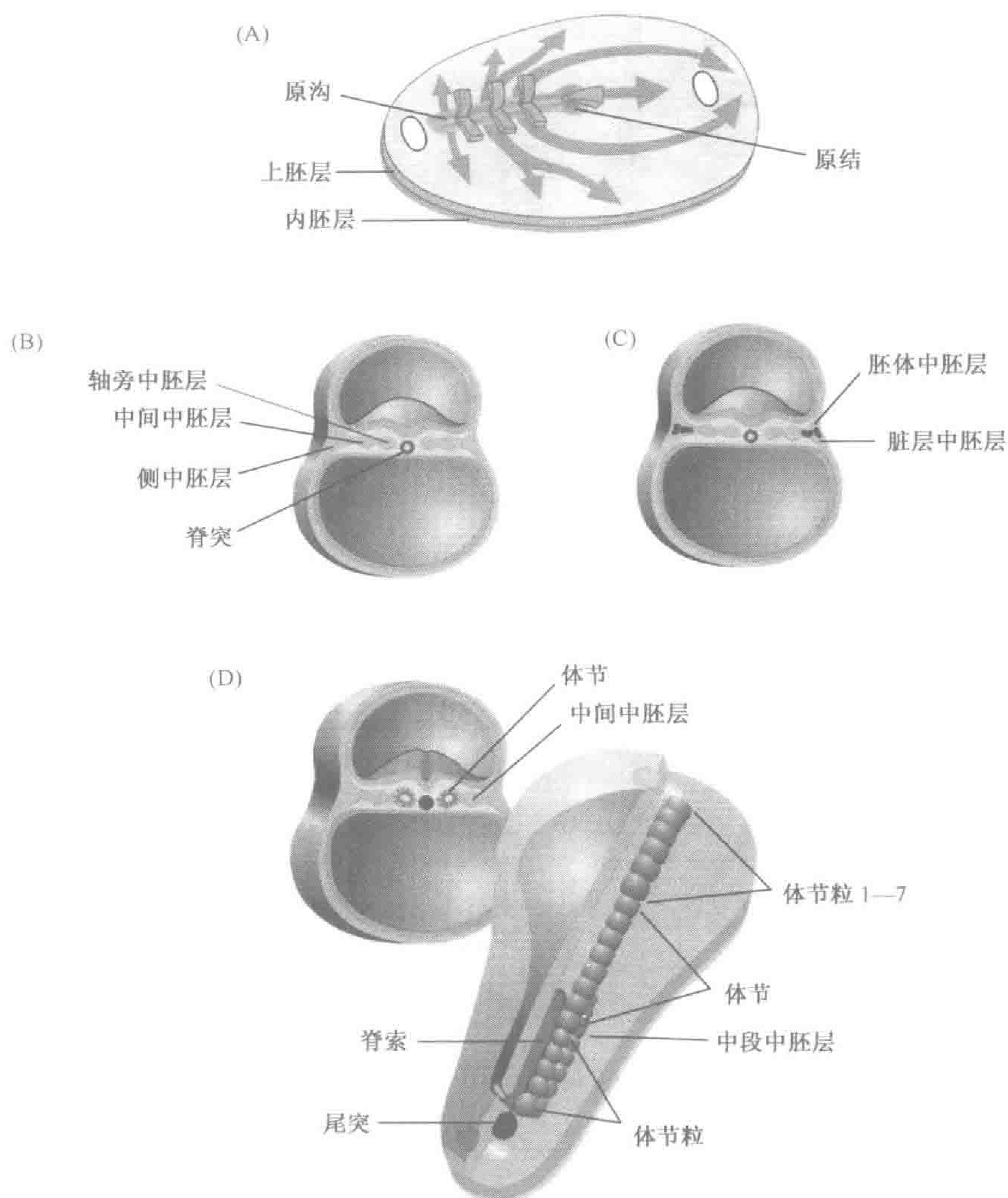


图 3.15　原肠胚形成期间进入的中胚层细胞的迁移途径和命运

(A) **早期中胚层迁移**。上胚层细胞经过原始结节迁移至头部形成前脊索板和头突，前脊索板是一个中胚层的致密团块，位于原始结节的头部，头突是一个高密度的中线管，将来形成一个实心棒，即脊索。上胚层细胞经过原沟迁移形成正中线侧面的侧中胚层。(B) **侧中胚层的早期分化**。与头突侧面紧密相连的中胚层形成圆柱状致密的轴旁中胚层。相邻不明显的圆柱状致密物是中段中胚层。侧中胚层的其余部分形成一个扁平层，即侧板中胚层。(C) **侧板中胚层**（LPM）**的分化**。LPM 内形成空泡分为两层：腹层即脏层中胚层，生成内脏器官的间皮覆盖物；背部胚体中胚层生成体壁的内衬和大部分真皮。(D) **体节粒形成**。轴旁中胚层形成一系列的圆形的螺纹样结构，即体节粒。除了体节粒 1～7（见正文），其余体节粒将会生成体节，即能形成身体节段结构的节段中胚层的块。在这个 21 天人类胚胎的示图中，六个中间体节粒已分化为体节。摘自 Larsen (2002)，经 Churchill Livingstone Publishers 允许。

- ▶ **近轴中胚层**（paraxial mesoderm，PM），一对柱形的致密结构，与脊索紧紧相邻并从侧面连接。近轴中胚层首先发育为一系列的螺纹状的结构，称为**体节粒**（somitomere），通过第三和第四周的发育，体节粒能形成一个头尾顺序。前七个体节粒最终形成脸、腭和喉的横纹肌，但其他的体节粒进一步发育为不连续的中胚层节段的块，称为**体节**（somite）（图 3.15）。颈、胸、腰和骶体节通过生成大部分轴骨架（包括脊柱）、随意肌和一部分皮肤真皮而建立躯体的节段结构。
- ▶ **中段中胚层**（intermediate mesoderm，IM）：一对不显著的圆柱形致密结构，就位于近轴中胚层的侧面。中段中胚层能发育为泌尿系统和一部分生殖系统。
- ▶ **侧板中胚层**（ lateral plate mesoderm，LPM）：侧面中胚层的残留部分，形成一个扁平层。从人类发育的第 17 天开始，侧板中胚层分裂为两层：
- ▶ **脏层中胚层**（splanchnopleuric mesoderm）：邻近内胚层那一层，能生成内脏器官的间皮覆盖物。
- ▶ **胚体壁中胚层**（somatopleuric mesoderm）：邻近外胚层的那一层，能生成体壁的内衬、一部分肢体和大部分皮肤的真皮。

3.8　神经发育

如上所述，原肠胚形成导致了胚胎一系列显著的变化，使其从两胚层结构转化为三胚层结构。现在胚胎的发育是规划将组织组成为成人体内的许多器官和系统的前体。我们集中在神经系统的早期发育作为器官发生的例子，因为它显示了分化的过程、模式形成和形态形成三者是如何精密协调的。

3.8.1　神经系统在外胚层被下面的中胚层诱导分化后发育

神经系统的发育标志着器官形成的起始，发生于人类发育的第三周末。启动事件是轴向中胚层对次级外胚层诱导（Wilson and Edlund，2001）。在轴向中胚层内，脊索前板和头索头部的细胞将信号转移至次级中胚层细胞导致其分化为神经上皮细胞厚板（**神经外胚层**，neurectoderm）。作为结果的**神经板**（neural plate）出现在人类发育的第 18 天，但接下来的两天，神经板生长十分迅速且改变了比例。

在第四周的开始阶段，称作**神经胚形成**（neurulation）的过程导致神经板转变为**神经管**（neural tube）——脑和脊髓的前体。神经板开始沿着中线腹侧变皱，形成一个凹窝，称为神经沟。这被认为是对来自邻近对面的脊索的诱导信号的反应引起的发育。厚的神经褶沿神经沟旋转交会于背侧，最初是在头尾轴的中点（图 3.16A）。神经管的关闭以拉链状和双向方式进行。在关闭不全的地方，有一个前神经孔和一个后神经孔。神经管偶尔地部分关闭不全而导致诸如脊柱裂疾病。

在神经胚形成期间，一群特殊的细胞沿着神经褶的侧边出现，与神经板分离，迁移至体内的许多特殊部位。这些高度多能的细胞群称为**神经嵴**（neural crest），可生成部分外周神经系统、黑素细胞、一些骨和肌肉、视网膜和其他结构（Garcia-Castro and Bonner-Fraser，1999；Knecht and Bonner-Fraser，2002；框 3.5）。

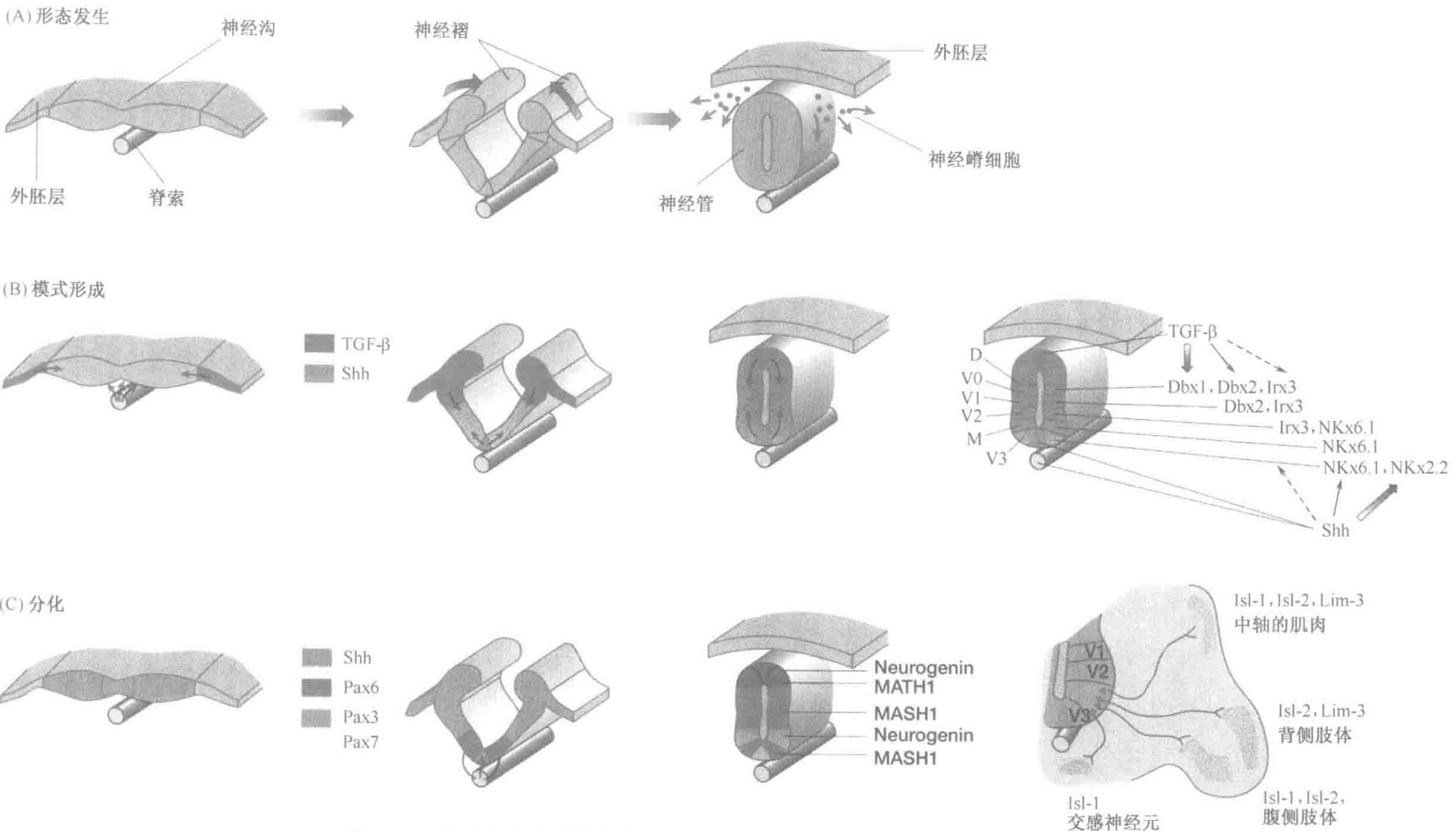


图 3.16 在器官发生的过程中，神经系统作为一个不同发育过程协调的例子

(A) 形态发生 (morpho genesis)。从扁平的神经板到封闭的神经管转变，是由于转折点的形成（此处细胞在顶部收缩）以及神经板边缘的外胚层细胞增殖推动两侧合在一起。当神经褶合在一起的时候，神经嵴细胞分层并开始迁移。(B) 模式形成 (pattern formation)。背腹模式的形成涉及来自脊索 (Sonic hedgehog) 和来自外胚层的分泌信号（开始是 BMP4，后来是 TGF-B 超家庭的其他成员）之间的竞争。当神经板折叠的时候，信号在神经管自身表达。结果是不同的转录因子在不同的区域被激活，导致在不同的区域不同的神经元亚型特化（D= 背部； V0、V1 等 = 腹部亚型； M= 运动神经元）。(C) 分化 (differentiation)。在发育中的神经系统中，控制神经发生的基因（原神经的和神经元发生的基因）也在沿着背腹轴的个别区域表达，很可能在 *Pax3*, *Pax6*, *Pax7* 基因的控制之下，这些基因的表达区部分由源自脊索的 Sonic hedgehog 信号确定。原神经的 neurogenin、*MATH1* 和 *MASH1* 的表达区域已在图中标示，而神经元发生的基因 *Notch*, *Delta* 和 *Serrate* 则没有。神经的发育以转录因子 NeuroD 的表达为特征，然后不同类型的神经元开始表达独特的转录因子群。LIM 同源结构域家族的转录因子通过调控突出轴突的行为，在确定亚神经元命运的时候尤其重要。

3.8.2　神经管的模式形成涉及沿着两个轴的基因的协同表达

神经板一旦形成，三个大的颅顶囊（将来的大脑）就可看见，将来可发育为脊髓的窄的尾节也同时可见。这一头尾向的极性反映神经诱导的区域特异性。来自轴向中胚层的信号包含位置信息，使次级外胚层形成轴的不同部分特异的神经组织。脊椎动物中这类信号的准确特性还不清楚（Stern，2002）。用非洲爪蟾建立一个较好的模型，其神经发育可被信号蛋白骨形态发生蛋白（BMP）家族中的部分成员抑制，神经诱导是被BMP的拮抗物激活的。然而在鸟类和哺乳类动物中该机制似乎更为复杂，是积极研究的领域。一般的神经形成信号似乎由中胚层释放，而中胚房适合诱导位于前面的神经板的形成。这一定是通过起源于胚胎尾区的其他信号变为“后部”的。头部的形成需要前部中胚层特异分泌的其他分子。

不论潜在的机制是什么，信号沿着轴激活不同的转录因子组合，而这些赋予细胞位置类别并调控其行为。在前脑和中脑，表达 Emx 和 Otx 家族转录因子。在后脑和脊髓，位置类别由 *Hox* 基因控制。在后脑，细胞的位置值似乎固定在神经板阶段，携带此位置信息的迁移神经嵴细胞把位置类别强加给周围的组织。相反，脊髓的位置类别是由来自周围近轴中胚层的信号所强加的。通过移植细胞到轴的不同部位可证明这一点。颅神经嵴细胞移动到新的位置时，它们依据其谱系来表现，即它们的行为与在原始位置时相同。另一方面，躯体神经嵴细胞依据它们的新位置来表现，即它们与其邻居的行为相同。

由于所有的细胞（不同种类的神经元和神经胶质）都沿着背腹轴发育，发育中的神经系统是模式形成和分化的一个好模型。这个过程由多级遗传调节子控制（Lumsden and Krumlauf，1996；Tanabe and Jessel ，1996；Eislen 1999；图 3.16B，C）。

背腹极通过来源于脊索和相邻外胚层的相反的几组信号产生。脊索分泌信号分子，它有腹侧化活性，而外胚层分泌几种转化生长因子（TGF）——β 超级家族，包括BMP4、BMP7 和一种称为 Dorsalin 的蛋白。当神经板开始折叠的时候，这些相同的信号开始在神经管自身的背区及腹区顶端——各自的基板和顶板表达。相反的信号对各种同源结构域类型转录因子的激活产生相反的影响。如图 3.16B 所示，这把神经管分为不连续的背腹区或转录因子结构域，后来会发育为不同种类的神经元的区域中心。例如，*Nkx6.1* 单独表达所确定的区域会成为运动神经元定位的区域，定位于神经管的腹侧第三部分。

3.8.3　神经元的分化涉及转录因子的组合行动

神经元不是统一在神经外胚层生成，而是限制于由**原神经基因**（proneural gene）（诸如神经生成蛋白、*MASH1* 和 *MATH1*，图 3.16C，编码 bHLH 家族转录因子）的表达所区分的特定区域。原神经基因表达赋予细胞形成成神经细胞的能力，但并不是所有的原神经细胞都有这种命运。相反，含有**神经发生基因**（neurogenic gene）（如 *Notch* 和 *Delta*）表达的细胞之间存在竞争，获得成功的细胞形成成神经细胞且抑制周围细胞的同样行为。因此，成神经细胞的形成有其精确的空间模式。这种模式的形成过程称为**侧抑制**（lateral inhibition）。然后神经元按照各自的神经系统背轴和腹轴的位置开始分化，由转录因子的

组合控制，就像前面章节讨论过的那样（Panchision and McKay，2002）。进一步的多样化由这些转录因子精细的表达模式所控制。例如，所有的运动神经元最先表达两种 LIM 同源结构域家族的转录因子：Islet-1 和 Islet-2。后来，只有那些将它们的轴突伸向腹侧肢体肌肉的运动神经元表达这些 LIM 转录因子，而其他运动神经元则不能。表达 Isl-1、Isl-2 和第三个转录因子 Lim-3 的神经元将它们的轴突伸向体壁的轴向肌肉。表达 Isl-2 和 Lim-1 的神经元将它们的轴突伸向背侧肢体肌肉，而那些仅表达 Isl-1 的神经元将它们的轴突伸向交感神经节（图 3.16C）。转录因子决定表达在轴突生长锥体上的受体的特定组合，因此决定生长的轴突对不同的物理和化学（化学引诱物质等）信号的反应（Tessier-Lavigne and Goodmans，1996）。一旦第一个轴突到达目标，其余的轴突通过沿着存在的轴突路径生长来发现自己的路径——这一过程称为自发性收缩。

3.9 发育途径的保守性

3.9.1 许多人类疾病是由于正常发育过程障碍所致

大多数引人注目的人类疾病都涉及由于分化、模式形成和形态发生的一般过程破坏所致的明显形态畸形。例如前脑无裂畸形是由于正常的前脑发育故障造成的，最严重的时候可发育成只有一只眼睛（独眼）或没鼻子的个体。正如其他疾病一样，表现型能受基因和环境因素双重影响。在一些情况下，特殊的突变导致畸形是非常明确的，如在编码信号蛋白“Sonic hedgehog”的 *SHH* 基因上突变。另外，畸形也可追找到环境因素的原因，如母体的饮食中摄入有限的胆固醇。通过性别决定机制可清晰的证实在发育的过程中基因和环境的相关功能，这些在框 3.9 中讨论。

框 3.9　性别决定：基因和发育的环境

人类和所有的哺乳动物一样为**二态性别**（sexually dimorphic），也就是分为雄性和雌性。将来发育为雄性还是雌性在受孕的时候就决定了，受孕的时候精子将 X 或 Y 染色体送入卵中，而卵中只有一种 X 染色体（Schafer and Goodfellow，1996）。仅有的例外发生于减数分裂发生错误的时候，此时配子中丢失或有过多的性染色体，导致非整倍体的个体出现（节 2.5.2）。尽管人类胚胎的性别在受孕的时候就已经确定了，但直到大约第 5 周的时候性别分化才开始。性别分化有两种形式，一种是**原始性征**（primary sexual characteristics）（生殖腺的发育和精子及卵之间的发育选择），另一种是**第二性征**（secondary sex characteristics）（泌尿系统和外生殖器的性别特殊结构）。原始性征的形成依赖于基因型，而第二性征的建立依赖于环境中的激素信号介导的信号。

男性的发育依靠 Y 染色体的存在或缺如，Y 染色体上有重要的称为 *SRY*（Y 染色体性别决定区）的男性决定基因。*SRY* 编码的转录因子能激活睾丸发育所需的下游基因。睾丸然后产生雄性第二性征发育所必需的性激素。很长一段时间，人们一直认为女性生殖腺的发育是一个‘缺欠状态’，且 *SRY* 基因足以将未分化的女性胚胎性腺转为男性分化。这种观点被几条证据支持——罕见的 XX 男性个体经常有一个携带 *SRY* 基因的 Y 染色体片段，易位在一条 X 染色体顶端。遗传为雌性的小鼠转入小鼠 *Sry* 基因即可发育为雄鼠。然而最近的研究表明 X 染色体上的基因和常染色体与卵巢发育的正调控有关。这些基因的过表达，包括 *DAX* 和 *WNT4A*，能使 XY 的个体女性化，即使他们具有功能的 *SRY* 基因。

框 3.9　性别决定：基因和发育的环境（续）

配子早期的发育受环境的影响比受基因型的影响要多。女性原始生殖细胞（PGC）放入睾丸中就会分化为精子，而男性原生殖细胞放入卵巢中即可分化为卵细胞。这可能反映了细胞周期的调控，PGC 进入睾丸中停止于减数分裂之前，而进入卵巢中则立即进行减数分裂。因此，由于没有信号阻止细胞周期，所以移植到生殖腺以外的体细胞组织中、任何性别的 PGC 开始分化为卵细胞。然而，在所有的这些不寻常的情况下，都没有产生有功能的配子。分化中止于相对较晚的阶段，大概是因为生殖细胞本身的基因型在配子的发育过程中也起到重要的作用。

不像原始性征的情况那样，好像女性的第二性征是一个缺欠状态。受 SRY 调控的基因之一是 *NR5A1*，编码称为类固醇生成因子 1（SF1）的转录因子。这个基因可以活化产生男性激素所需的基因，包括 *HSD1783*（编码 hydroxysteroid-17-β-脱氢酶 3，为睾酮合成所需要）和 *AMH*（该基因编码抗苗勒氏激素）。上述两种基因在男性的泌尿生殖系统分化中起重要作用。例如 *AMH* 能使苗勒氏管（在女性中会发育为输卵管和子宫）破裂。抑制这种激素产物的产生、分配、消失或感知的突变能使 XY 基因型的个体女性化。例如，雄激素不敏感综合征是由于睾酮受体缺陷所致，这能阻止身体对激素的反应，即使产生的激素在正常水平。患这种疾病的 XY 基因型个体表面上是一个正常女性，但由于 SRY 和 AMH 的作用，他们具有没有下降的睾丸而不是卵巢，他们也缺乏子宫和输卵管。导致女性中男性激素过量生产的突变会产生相反的影响，例如 XX 基因型的个体男性化。偶尔，这种现象发生于发育中的男/女双胞胎中，女性的胎儿暴露于她兄弟的雄激素中。CYP19 酶将雄激素转化为雌激素导致男性的女性化，而那些降低或抑制 CYP19 酶的活性的因素将导致女性的男性化。

一些发育的畸形可归因于药物使用（已知化学物质是引起畸形形成的因素，如酒精、某些的抗生素、镇静类药物、维甲酸和非法的毒品如可卡因和海洛因），更多的发育畸形是特殊的基因突变所引起。这一点在本章的开始就已经介绍了，一些最重要的发育基因本质上是调节器，编码转录因子或信号通路的成分。编码细胞、细胞外基质甚至代谢酶的结构组件的基因，在发育和揭示重要的疾病表现型上发挥重要功能。表 3.8 提供了一些例子。

表 3.8　人类发育基因的选择和特殊的疾病表现型相关。几个主要种类的基因产物被考虑：信号蛋白、受体、转录因子、结构蛋白和酶

基　因	产物和正常功能	相关疾病
分泌的信号蛋白		
SHH	Sonic hedgehog 负责神经管、体节、肠、和肢芽模式的信号蛋白	Holoprosencephaly，发育中前脑没有分为左右半球。严重的病倒只有一个脑室和独眼，在很轻的病例中可能表现为只有一颗中切牙
EDN3	内皮素 3，调节血管收缩的肽类激素，在发育中神经嵴衍生物的分化需要	**先天无神经节性巨结肠病**（Hirschsprung）一种神经嵴分化的缺陷，其肠道神经节没有形成，由于缺少蠕动形成巨结肠（慢性便秘和肠梗阻）
GH1	生长激素 1，由垂体分泌的多肽激素，负责调节生长	垂体性侏儒症可在儿童时期通过增加纯化的或重组的生长激素进行治疗的一种侏儒症

续表

基　因	产物和正常功能	相关疾病
LEFT8	Lefty2，结节相关信号蛋白，在早期胚胎的左手侧表达，帮助建立左右不对称	偏侧性缺陷——逆位、异构或内脏异位，由左右轴特化的失败引起
受体		
FGFR3	成纤维细胞生长因子受体，在软骨和神经系统中表达特别强。在骨发育中起主要作用	软骨发育不全，短肢侏儒最常见的形式 Crouzon 综合征和颅缝早闭，严重的颅面畸形
EDNRB	内皮素 B 受体，发现于神经嵴细胞，在神经嵴细胞分化为黑素细胞和肠内肠神经节时需要	先天无神经节性巨结肠病（EDN3）
KIT	KIT 是一种酪氨酸激酶受体，最初发现是猫肉瘤病毒中的癌基因因而如此称呼。该受体广泛表达，但在血细胞系、黑素细胞和生殖细胞发育中有特别重要的功能。	白化病，以先天性的皮肤和头发白斑为特征，由黑素细胞增殖障碍引起。
GHR	生长激素受体，转导生长激素信号（见上文）	Laron 侏儒症，是一种血清生长激素水平正常的侏儒症，对生长激素治疗不敏感。也被称为生长激素不敏感综合征
转录因子		
HOXD13	与模式形成有关的转录因子。沿头尾轴在肢体及在发育中传递位置信息。	由于细胞类型的特化错误和随后形成的肢体远侧区域骨结构异常导致的多指（趾）畸形［额外融合指（趾）］
PAX6	眼睛发育过程中具多种功能的转录因子	从轻微的瞳孔异位（瞳孔偏位）到虹膜缺失（部分或全部的虹膜缺失，合并晶状体、前房畸形和角膜变性）的眼缺陷
TBX5	特异表达于发育中胚胎的前肢区的转录因子，在肢体特性的建立中发挥重要作用（臂对腿发育）	心手综合征，一种上肢疾病，也影响心脏的发育
SRY	在男性胚胎生殖嵴表达的一种转录因子，在启动男性性别分化的时候需要，发现于 Y 染色体上	男性性别逆转（XY 个体外观为女性表现）常与完全的性腺发育不全相关
结构蛋白		
COL6A1	是Ⅵ型胶原的 α 亚单位，为组织提供结构强度的细胞外基质微纤维的主要成分	Bethlem 肌病，表现为关节挛缩（特别是肘和踝），由缺乏Ⅵ型胶原微纤维引起。Down 综合征中过度表达可能导致心脏缺陷
ELN	弹性蛋白，细胞外基质的一种蛋白，可使组织变形后恢复为原来形状	皮肤松弛症，由于弹性蛋白异常或功能障碍导致皮肤永久变形引起松垂。 主动脉狭窄，由主动脉变形引起
LAMA3	层粘连蛋白 5 的 α3 亚单位，是皮肤基底细胞的一种成分，在角质细胞分化过程中起重要作用	结合性大疱性表皮松解，基底细胞脱离基底膜所致的一种皮肤起泡疾病。
USH2A	眼和内耳发育所需的一种细胞外基质蛋白	Ⅱ型 Usher 综合征，以出生时就有严重的耳聋和青春期晚期色素性视网膜炎为特征

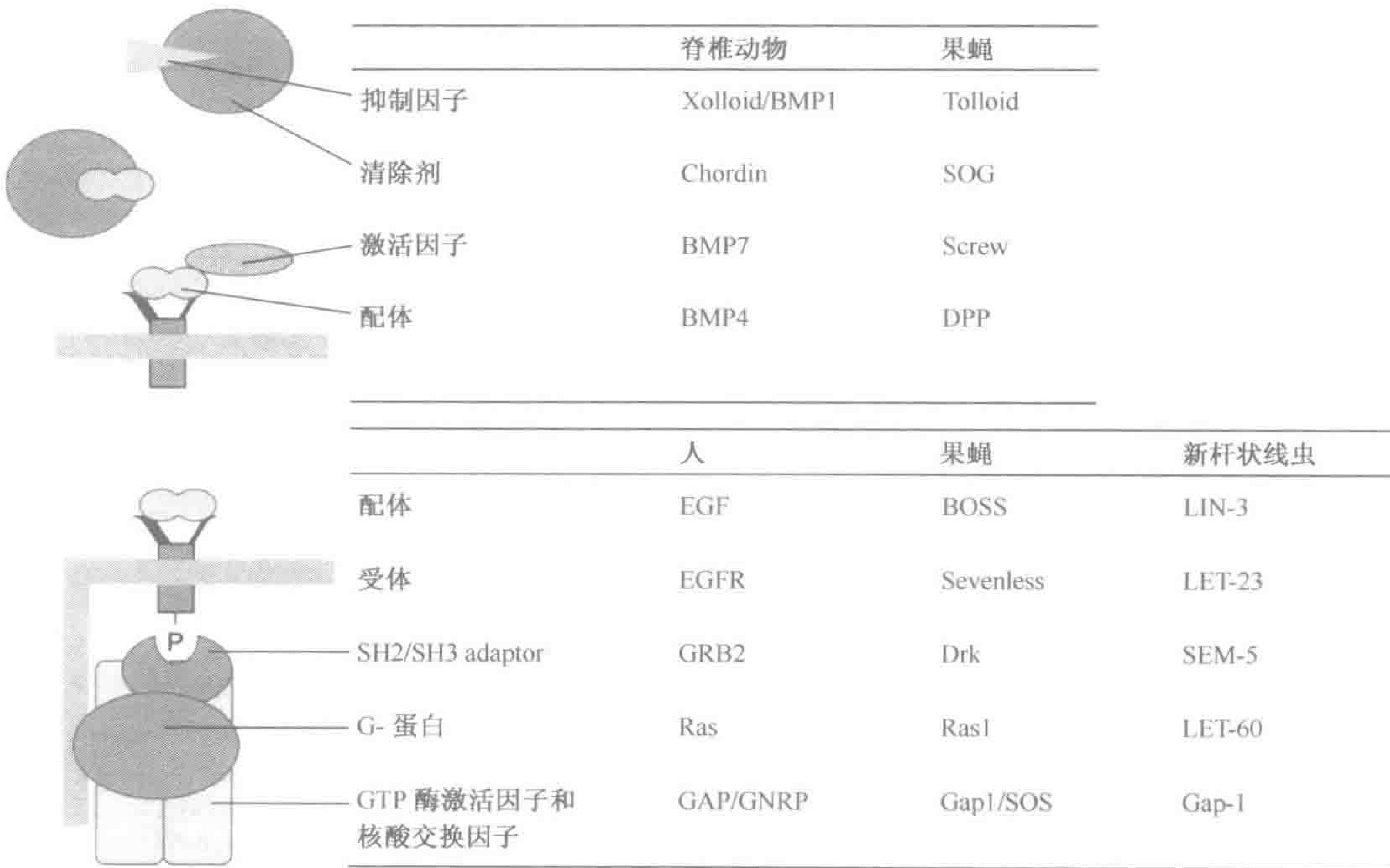
续表

基 因	产物和正常功能	相关疾病
酶		
WRN	解旋酶，可能与双链DNA断裂的修复有关	Werner综合征，一种早熟性老化疾病
CYP11B1	类固醇11-β-羟化酶，为醛固酮的合成所需。（一种在肾脏作用的激素，调节盐水平衡）	先天性肾上腺增生，表现为儿童快速生长但骨延长早熟终止，导致成人身材矮小
HSD17B3	17-β-羟化类固醇脱氢酶，睾丸酮合成所需	男性假两性畸形，男孩出生时有明显的女性外观，但在青春期时男性化
EXT1	糖基转移酶，在合成硫酸乙酰肝素时需要，是细胞外基质的一种重要成分	遗传性多发性外生骨疣，是一种靠近骨端的软骨结块生长的疾病，特别是在肢体，但偶尔也出现于肋骨和肩

3.9.2 在单基因水平和整体通路水平，发育过程都是高度保守的

在动物中，被证实导致发育疾病的基因经常有显著程度的进化保守性。这种保守性不仅适用于基因，而且适用于基因所涉及的整个通路（Pires-daSilva and Sommer，2003；图3.17）。

保守的分子通路在亲缘关系相距很远的物种中经常用于相似的进程。例如，如前面



	脊椎动物	果蝇
抑制因子	Xolloid/BMP1	Tolloid
清除剂	Chordin	SOG
激活因子	BMP7	Screw
配体	BMP4	DPP

	人	果蝇	新杆状线虫
配体	EGF	BOSS	LIN-3
受体	EGFR	Sevenless	LET-23
SH2/SH3 adaptor	GRB2	Drk	SEM-5
G-蛋白	Ras	Ras1	LET-60
GTP酶激活因子和核酸交换因子	GAP/GNRP	Gap1/SOS	Gap-1

图3.17 发育通路的进化保守性

（A）在果蝇和脊椎动物中，BMP4/Chordin是神经诱导的基础。

（B）生长因子信号通路在脊椎动物、蝇和蠕虫中有多种功能。

简述的那样，非洲爪蟾的神经诱导涉及 BMP4（BMP4 倾向于腹面和侧面的发育方向）和倾向背侧发育的因子（诸如 Chordin，Noggin 和 Follistatin）之间的相互对抗。BMP4 和 Chordin 在果蝇中的种间同源基因是分别称为 Decapentaplegic（Dpp）和 Short gastrulation（Sog）的蛋白质。引人注意的是，这些蛋白质在果蝇的神经系统形成的过程中起相同的作用。实际上关系更为密切。果蝇的 Dpp 的活性被蛋白质 Tolloid（Tol）增强，而 Tol 能降解 Sog。非洲爪蟾的种间同源基因 Xolloid 可增强 Tol 的功能，斑马鱼的 BMP1 也有相同的功能，两者都可抑制 Chordin 发挥作用。非洲爪蟾 BMP7 和果蝇蛋白 Screw 之间也有交叉，对 BMP4/Dpp 的活性来说附属蛋白是必需的。

另一种情况，在不同的物种中，相同的发育通路被用于极为不同的目的。在哺乳动物中，表皮生长因子（EGF）用于提高表皮细胞的增殖。这种生长因子可与 EGF 受体（是一种受体酪氨酸激酶）结合且能启动 Ras-Raf-MAP 激酶级联反应（节 3.2.1）。在果蝇眼部的发育过程中，这种相同的发育通路被用于提高八个光感受器细胞类型之一的分化。而在线虫的发育过程中，这种相同的发育通路被用于刺激外阴细胞的分裂和分化。

进化保守性的最好例子是同源框含有的基因，该基因出现在所有的动物中且行使相似的功能。在模式形成过程中这些基因的基本功能可被来自不同物种但能互相替代的种间同源基因的能力所证实。例如，将人类的 *HOX* 和 *OTX* 基因引入果蝇种间同源基因突变株，可引起突变表型的完全恢复。

然而，即使在那些亲缘很近的物种间，也可以有显著的区别。考虑到人和大鼠之间存在广泛相似性，你或许会对原肠胚形成过程有如此多的不同感到奇怪（图 3.14）。实际上脊椎动物胚胎在原肠胚形成之前有很大的不同，这反映了营养采集的不同策略。性别决定机制也是很多样的（Morrish and Sinclair，2002）。并不是所有的动物采用人类的 XY 性别决定模式（Graves，2002），许多爬行动物应用异形性染色体，完全依靠环境的温度来特定胚胎的性别。

（邱广斌 译）

进一步阅读

Arias AM, Stewart A (2002) *Molecular Principles of Animal Development.* Oxford University Press, Oxford.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P (2002) *Molecular Biology of the Cell*, 4th Edn. Garland Science, New York.

Gilbert SF (2003) *Developmental Biology*, 7th Edn. Sinauer Associates, Sunderland.

Twyman RM (2001) *Instant Notes in Developmental Biology.* BIOS Scientific Publishers, Oxford.

Wolpert L (2002) *Principles of Development*, 2nd Edn. Oxford University Press, Oxford.

参考文献

Bainter JJ, Boos A, Kroll KL (2001) Neural induction takes a transcriptional twist. *Dev. Dynamics* **222**, 315–327.

Beddington RSP, Robertson EJ (1999) Axis development and early asymmetry in mammals. *Cell* **96**, 195–209.

Bjornson CRR, Rietz RL, Reynolds BA *et al.* (1999) Turning brain into blood: a hematopoietic fate adopted by adult neural stem cells in vivo. *Science* **283**, 354–357.

Bokel C, Brown NH (2002) Integrins in development: Moving on,

responding to, and sticking to the extracellular matrix. *Dev. Cell* **3**, 311–321.

Bourne HR (1997) How receptors talk to trimeric G proteins. *Curr. Opin. Cell Biol.* **9**, 134–142.

Burke AC (2000) *Hox* genes and the global patterning of the somitic mesoderm. *Curr. Top. Dev. Biol.* **47**, 155–181.

Campisi J (1996) Replicative senescence: an old live's tale? *Cell* **84**, 497–500.

Capdevila J, Vogan KJ, Tabin CJ, Belmonte JCI (2000) Mechanisms of left-right determination in vertebrates. *Cell* **101**, 9–21.

Clapham DE (1995) Calcium signaling. *Cell* **80**, 259–268.

Clarke DL, Johansson CB, Wilbertz J *et al.* (2000) Generalized potential of adult neural stem cells. *Science* **288**, 1660–1663.

Cunningham BA (1995) Cell adhesion molecules as morphoregulators. *Curr. Opin. Cell Biol.* **7**, 628–633.

Daley GQ (2002) Prospects for stem cell therapeutics: myths and medicines. *Curr. Opin. Genet. Dev.* **12**, 607–613.

Davis AP, Witte DP, Hsieh-Li HM *et al.* (1995) Absence of radius and ulna in mice lacking *Hoxa-11* and *Hoxd-11*. *Nature* **375**, 791–795.

Donovan PJ, Gearhart J (2001) The end of the beginning for pluripotent stem cells. *Nature* **414**, 92–97.

Dustin ML (2002) Shmoos, rafts, and uropods – The many facets of cell polarity. *Cell* **110**, 13–18.

Eislen JS (1999) Patterning motoneurons in the vertebrate nervous system. *Trends Neurosci.* **22**, 321–326.

Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156.

Garcia-Castro M, Bonner-Fraser M (1999) Induction and differentiation of the neural crest. *Curr. Opin. Cell Biol.* **11**, 695–698.

Giancotti FG, Ruoslahti E (1999) Transduction – Integrin signalling. *Science* **285**, 1028–1032.

Graves JAM (2002) The rise and fall of *SRY*. *Trends Genet.* **18**, 259–264.

Gumbiner BM (1996) Cell adhesion: the molecular basis of tissue architecture and morphogenesis. *Cell* **84**, 345–357.

Harland R (2000) Neural induction. *Curr. Opin. Genet. Dev.* **10**, 357–362.

Hogan BLM (1999) Morphogenesis. *Cell* **96**, 225–233.

Hou SX, Zheng ZY, Chen X *et al.* (2002) The JAK/STAT pathway in model organisms: Emerging roles in cell movement. *Dev. Cell* **3**, 765–778.

Houslay MD, Milligan G (1997) Tailoring cAMP-signalling responses through isoform multiplicity. *Trends Biochem. Sci.* **22**, 217–224.

Humphries MJ, Newham P (1998) The structure of cell-adhesion molecules. *Trends Cell. Biol.* **8**, 78–83.

Hynes RO (2002) Integrins: Bidirectional, allosteric signaling machines. *Cell* **110**, 673–687.

Irvine KD, Rauskolb C (2001) Boundaries in development: formation and function. *Annu. Rev. Cell Dev. Biol.* **17**, 189–214.

Kim H, Schagat T (1996) Neuroblasts: a model for the asymmetric division of cells. *Trends Genet.* **13**, 33–39.

Knecht AK, Bonner-Fraser M (2002) Induction of the neural crest: A multigene process. *Nature Rev. Genet.* **3**, 453–461.

Krumlauf R (1994) *Hox* genes in vertebrate development. *Cell* **78**, 191–201.

Kuo CJ, Conley PB, Chen L *et al.* (1992) A transcriptional hierarchy involved in mammalian cell type specification *Nature* **355**, 457–461.

Larsen W (2002) *Human Embryology*, 3rd Edn. Churchill Livingstone, New York.

Lawrence PA, Morata G (1994) Homeobox genes: their function in *Drosophila* segmentation and pattern formation. *Cell* **78**, 181–189.

Le Mouellic H, Lallemand Y, Brulet P (1992) Homeosis in the mouse induced by a null mutation in the *Hox 3.1* gene. *Cell* **69**, 251–264.

Lu CC, Brennan J, Robertson EJ (2001) From fertilization to gastrulation: axis formation in the mouse embryo. *Curr. Opin. Genet. Dev.* **11**, 384–392.

Lubarsky B, Krasnow MA (2003) Tube morphogenesis: Making and shaping biological tubes. *Cell* **112**, 19–28.

Lumsden A, Krumlauf R (1996) Patterning the vertebrate neuraxis. *Science* **274**, 1109–1115.

Manouvrier-Hanu S, Holder-Espinasse M, Lyonnet S (1999) Genetics of limb anomalies in humans. *Trends Genet.* **15**, 409–417.

Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl Acad. Sci. USA* **78**, 7634–7638.

Mathis L, Nicolas J-F (2002) Cellular patterning of the vertebrate embryo. *Trends Genet.* **18**, 627–635.

McNeill H (2000) Sticking together and sorting things out: adhesion as a force in development. *Nature Rev. Genet.* **1**, 100–108.

Medvinsky A, Smith A (2003) Stem cells: Fusion brings down barriers. *Nature* **422**, 823–825.

Morata G (1993) Homeotic genes of *Drosophila*. *Curr. Opin. Genet. Dev.* **3**, 606–613.

Morrish BC, Sinclair AH (2002) Vertebrate sex determination: many means to an end. *Reproduction* **124**, 447–457.

Morrison SJ, Uchida N, Weissman IL (1995) The biology of hematopoietic stem cells. *Ann. Rev. Cell Biol.* **11**, 35–71.

Myers DC, Sepich DS, Solnica-Krezel L (2002) Convergence and extension in vertebrate gastrulae: cell movements according to or in search of identity? *Trends Genet.* **18**, 447–455.

Narasimha M, Leptin M (2000) Cell movements during gastrulation: come in and be induced. *Trends Cell Biol.* **10**, 169–172.

Ng JK, Tamura K, Buscher D *et al.* (1999) Molecular and cellular basis of pattern formation during vertebrate limb development. *Curr. Top. Dev. Biol.* **41**, 37–66.

Niswander L (2003) Pattern formation: Old models out on a limb. *Nature Rev. Genet.* **4**, 133–143.

Orlic D, Kajstura J, Chimenti S, Jakoniuk I *et al.* (2001) Bone marrow cells regenerate infarcted myocardium. *Nature* **410**, 701–705.

Panchision DM, McKay RDG (2002) The control of neural stem cells by morphogenetic signals. *Curr. Opin. Genet. Dev.* **12**, 478–487.

Peifer M, McEwen DG (2002) The ballet of morphogenesis: Unveiling the hidden choreographers. *Cell* **109**, 271–274.

Pellegrini S, Dusanter-Fourt I (1997) The structure, regulation and function of the Janus kinases (JAKs) and the signal transducers and activators of transcription (STATs). *Eur. J. Biochem.* **248**, 615–633.

Perez-Moreno M, Jamora C, Fuchs E (2003) Sticky business: Orchestrating cellular signals at adherens junctions. *Cell* **112**, 535–548.

Petersen BE, Bowen WC, Patrene KD *et al.* (1999) Bone marrow as a potential source of hepatic oval cells. *Science* **284**, 1168–1170.

Pires-daSilva A, Sommer RJ (2003) The evolution of signalling pathways in animal development. *Nature Rev. Genet.* **4**, 39–49.

Robinson MJ, Cobb MH (1977) Mitogen activated kinase pathways. *Curr. Opin. Cell Biol.* **9**, 180–186.

Saucedo LJ, Edgar BA (2002) Why size matters: altering cell size. *Curr. Opin. Genet. Dev.* **12**, 565–571.

Schafer AJ, Goodfellow PN (1996) Sex determination in humans. *Bioessays* **18**, 955–963.

Schoenwolf GC, Smith JL (1990) Mechanisms of neurulation: traditional viewpoint and recent advances. *Development* **109**, 243–270.

Schwabe JWR, Rodriguez-Esteban C, Belmonte JCI (1998) Limbs are moving: Where are they going? *Trends Genet.* **14**, 229–235.

Selleck SB (2000) Proteoglycans and pattern formation: sugar biochemistry meets developmental genetics. *Trends Genet.* **16**, 206–212.

Shamblott MJ, Axelman J, Wang SP, Bugg EM *et al.* (1998)

Derivation of pluripotent stem cells from cultured human primordial germ cells. *Proc. Natl Acad. Sci. USA* **95**, 13726–13731.

Shih CC, Mamelak A, LeBon T, Forman SJ (2002) Hematopoietic potential of neural stem cells. *Nature Med.* **8**, 535–536.

Speigel S, Foster D, Kolesnick R (1996) Signal transduction through lipid second messengers. *Curr. Opin. Cell Biol.* **8**, 159–167.

Stacey G, Doyle A (2000) Cell banks: A service to animal cell technology. In: *Encyclopedia of Animal Cell Technology* (ed Spier R). Wiley, New York, pp. 293–320.

Steinberg MS, McNutt PM (1999) Cadherins and their connections: adhesion junctions have broader functions. *Curr. Opin. Cell Biol.* **11**, 554–560.

Stern CD (2002) Induction and initial patterning of the nervous system – the chick embryo enters the scene. *Curr. Opin. Genet. Dev.* **12**, 447–451.

Streuli C (1999) Extracellular matrix remodelling and cellular differentiation. *Curr. Opin. Cell Biol.* **11**, 634–640.

Su TT, O'Farrell PH (1998) Size control: cell proliferation does not equal growth. *Curr. Biol.* **19**, R687–R689.

Tabin CJ, Vogan KJ (2003) A two-cilia model for vertebrate left-right axis specification. *Genes Dev.* **17**, 1–6.

Tanabe Y, Jessel TM (1996) Diversity and pattern in the developing spinal cord. *Science* **274**, 1115–1123.

Tessier-Lavigne M, Goodmans CS (1996) The molecular biology of axon guidance. *Science* **274**, 1123–1133.

Thomson JA, Itskovitz-Elder J, Shapiro SS *et al.* (1998) Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147.

Tsai MJ, O'Malley BW (1994) Molecular mechanisms of action of steroid/thyroid receptor superfamily members. *Annu. Rev. Biochem.* **63**, 451–486.

Twyman RM (2001) Signal transduction in development. In: *Instant Notes in Developmental Biology*. BIOS Scientific Publishers, Oxford.

Vaux DL, Korsmeyer SJ (1999) Cell death in development. *Cell* **96**, 245–254.

Verfaile CM (2002) Adult stem cells: assessing the case for pluripotency. *Trends Cell. Biol.* **12**, 502–508.

Wassarman PM (1999) Mammalian fertilization: Molecular aspects of gamete adhesion, exocytosis, and fusion. *Cell* **96**, 175–183.

Weissman IL (2000) Translating stem and progenitor cell biology to the clinic: barriers and opportunities. *Science* **287**, 1442–1446.

Wilson SI, Edlund T (2001) Neural induction: toward a unifying mechanism. *Nature Neurosci.* **4(Suppl)**, 1161–1168.

Wolpert L (1996) One hundred years of positional information. *Trends Genet.* **12**, 359–364.

Wylie C (2000) Germ cells. *Curr. Opin. Genet. Dev.* **10**, 410–413.

Zernicka-Goetz M (2002) Patterning the embryo: the first spatial decisions in the life of a mouse. *Development* **129**, 815–829.

第4章　系谱及群体中的基因

本章内容

4.1　单基因与多因子遗传

最简单的遗传性状是指其存在与否依赖于单一**基因座**（locus）上的**基因型**（genotype）的性状。这并非是说该性状本身仅由一对基因决定：任何人类性状的表达可能需要诸多基因和环境因素。然而，在一定人类正常遗传的和环境背景下，有时某一基因座上特定的基因型对其性状的表达是必要和充分的。这样的性状被称为**孟德尔式**（Mendelian）。孟德尔式性状可以通过特征性的系谱方式来识别（节 4.2）。在人类已知有 6000 多种孟德尔式性状。正如在"我们智能化使用网络之前"中描述的那样，获得任何这类病理或非病理性状相关信息的主要起点就是 OMIM 数据库（http://www.ncbi.nlm. nin. gov/Omim/）。

大多数人类性状是由一个以上基因座的基因决定的。某一性状与其主要基因作用的关系越远，就越不可能表现为简单的孟德尔系谱方式。DNA 序列的变异实际上总完全是孟德尔式的，这也是其作为遗传标记的主要优点（节 13.2）。蛋白质变异体（电泳迁移率或酶活性不同）一般是孟德尔式的，但由于有转录后的修饰作用，其可能取决于多

个基因座（节 1.5.3）。导致出生缺陷的发育通路的异常很可能涉及多因素的复杂平衡关系。因此，常见的出生缺陷（腭裂、脊柱裂、先心病等）极少是孟德尔式的。行为性状如 IQ 测试结果、精神分裂症也很少可能是孟德尔式的，但它们在不同程度上是遗传决定的。

非孟德尔遗传性状可能取决于二、三个或更多的基因座，也受不同程度环境因素的影响。此处，我们用**“多因子”**（multifactorial）作为总括词来概括所有这些可能性。更为特殊的是，遗传决定因素可能涉及少数基因座［**寡基因的**（oligogenic）］，或者每个起微小效应的许多基因座［**多基因**（polygenic）］，或者可能存在具有多基因背景的一个主要基因座。对于**双歧性状**（dichotomous character）（或有或无的性状，如多指）来说，这些潜在的基因座被假定为**易感性基因**（susceptibility gene），而对于**数量性状**（quantitative character）或**连续性状**（continuous character）（身高、体重等）来说，它们被视为**数量性状遗传基因座**（QTL）。该类性状中的任一个均可以在家系中传递，但其系谱方式不符合孟德尔式，且不可能简单的分析。

4.2　孟德尔式系谱类型

4.2.1　显性和隐性是性状的而不是基因的特性

某一性状如果在**杂合子**（heterozygote）中表现，就是**显性**（dominant），反之，则为**隐性**（recessive）。值得注意的是，显性、隐性是性状的而不是基因的特性。因此，镰形细胞贫血是隐性的，因为只有在 HbS 为**纯合子**（homozygote）时才表现出症状。但镰形性状是 HbS 杂合子的表型，故为显性的。大多数的人类显性疾病仅发生在杂合子，有时有两个杂合子患者婚配所生的纯合子，通常病情更为严重。例如软骨发育不全（短肢侏儒，MIM 100800）和Ⅰ型 Waardenburg 综合征（伴有色素异常的耳聋；MIM 193500）。无论如何，我们称软骨发育不全和Ⅰ型 Waardenburg 综合征为显性，是因为这些词汇描述在杂合子中所见的**表型**（phenotype）。在实验的有机体中不存在这种不确定性，当杂合分子具有中间表型，对纯合子与杂合子是无法分辨的疾病——例如 Huntington 病（成人发病的进行性神经退化，MIM 143100）保留“显性”时，遗传学家倾向于“半显性”一词。对显性这个问题，Wilkie（1994）已作了很好的综述。对 X、Y 染色体上的基因座来讲，男性是**半合子**（hemizygou），其每个基因仅有单一拷贝，因此，X 或 Y 连锁性状的显性或隐性的问题在男性中不会发生。

4.2.2　有五种基本的孟德尔式系谱方式

图 4.1 给出用于系谱绘制的符号，框 4.1 总结了每种方式的主要因素。孟德尔性状由**常染色体**（autosome chromosome）或 X、Y **性染色体**（sex chromosome）上的基因座决定。在两种性别中的常染色体性状以及在女性中的 X 连锁性状可能是显性的或隐性的。没有人在遗传上具有两个不同的 Y 染色体（在罕见的 XYY 核型的男性，两条 Y 染色体是重复的）。因此，共有五种典型的孟德尔式系谱方式（图 4.2）。特殊考虑用于如下文所述的 X 连锁或 Y 连锁疾病，因此，在实际上重要的孟德尔遗传方式是常染色体显性的、常染色体隐性的和 X 连锁的（显性的或隐性的）。这些基本的系谱方式将遇

到各种复杂的情况，将在节 4.3（下文）中讨论和图 4.5 中来举例说明。



图 4.1　用于系谱中的主要符号

世代通常以罗码数字表示，每代中各成员以阿拉伯数字编号，Ⅲ-7 或 Ⅲ$_7$ 指在第Ⅲ代中左起第七个人（如果没有其他特殊编号）。箭头可以用来表示先证者，通过其来确证该家系。

框 4.1　孟德尔遗传方式的特征

常染色体显性遗传（图 4.2A）：

受累者通常至少有一个受累的双亲（例外见图 4.4）；

两性均受累；

两性均可传递致病基因；

患者与正常人婚配的子女有 50%概率患病（假设患者是杂合子，对于罕见的疾病通常是真实的）。

常染色体隐性遗传（图 4.2B）：

患者通常生于非受累的双亲；

患者的双亲通常是无症状的携带者；

双亲近亲婚配发病率增高，两性均受累；

已生一个患儿后，每再生一孩有 25%的受累概率（假设双亲是表型正常的携带者）。

X 连锁隐性遗传（图 4.2C）：

主要累及男性；

受累的男性通常生于非受累的双亲；母亲是表型正常无症状的携带者，而可能有受累的男性亲属；

如果父亲为受累者且母亲是携带者，或者偶发于 X 染色体非随机的失活，那么女性可能受累（节 4.2.2）；

在系谱中无男性-男性的传递（但受累男性和女性携带者的婚配会造成由男性传递给男性的表象，图 4.5G）。

X 连锁显性遗传（图 4.2D）：

两性均受累，但女性多于男性；

通常女性比男性受累程度更轻且变异更大；

受累女性的孩子无论男女均有 50%受累几率；受累男性的女儿均受累，儿子则无一受累。

Y 连锁遗传（图 4.2E）：

仅男性受累。

受累男性总有一个受累的父亲（除非有一新的突变）。

受累男性所有的儿子均受累。

X-失活（莱昂作用）混淆了显性和隐性 X-连锁疾病之间的区别

隐性 X-连锁疾病的携带者通常会表现出这种疾病的某些体征，而与受累的男性相比，显性 X-连锁疾病的杂合子通常症状更轻、变异更大。这是 **X-**失活（X-inactivation）

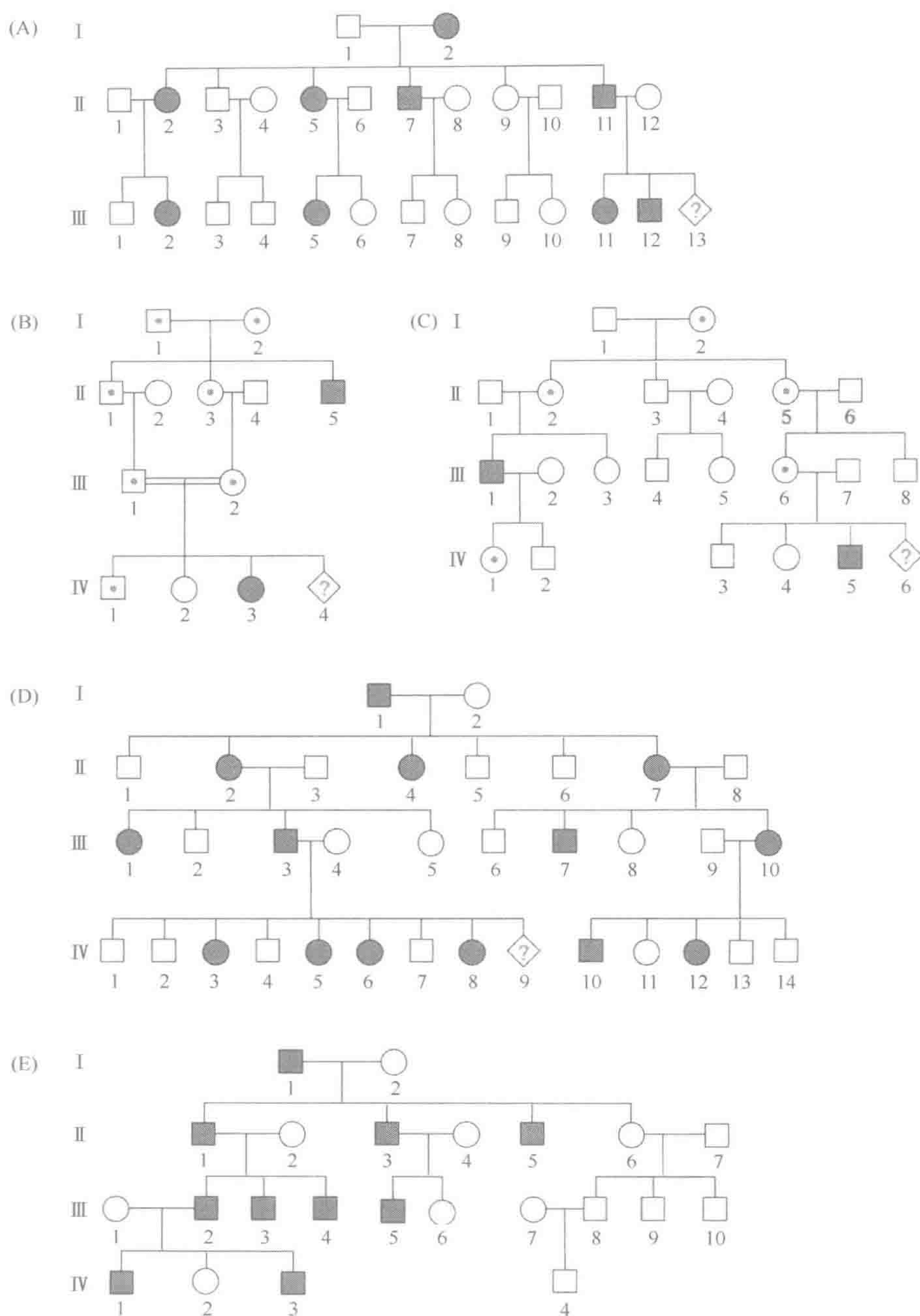


图 4.2 基本的孟德尔遗传方式

(A) 常染色体显性的。(B) 常染色体隐性的。(C) X连锁隐性的。(D) X连锁显性的。(E) Y连锁的。用问号标记的个体的发病风险是 (A) 1/2 (B) 1/4 (C) 1/2 男性或 1/4 全部后代; (D) 男性发病率低可忽略，女性 100%发病。对这些基本方式的详情见节 4.3 和图 4.5。

的结果。正如节 10.5.6 中所述，哺乳动物通过在每个体细胞中永久地使所有的 X 染色体（除了一条外）失活来补偿男女两性中的 X 染色体的不等数目。在每个细胞中，XY

男性保持其仅有的一条 X 染色体有活性，XX 女性使一条 X 染色体（随机选择）失活。失活发生于胚胎早期，一旦细胞选择了哪一条 X 染色体失活，那么这种选择被克隆性地传递给其所有的子细胞。

一个 X 连锁疾病（显性的或隐性的）的女性杂合子是一个嵌合体（节 4.3.6）。每个细胞表达正常或异常的等位基因，但二者不是同时表达。在表型依赖于血液循环产物的情况下，正如在血友病（凝血障碍，MIM 306700，306900）中，正常和异常细胞之间存在均分的效应。女性携带者具有中间表型，临床上通常不发病，但生化检查上是异常的。其表型是个别细胞的局限化的特征，正如在少汗型外胚层发育不良（无汗腺，异常的牙齿、头发；MIM 305100）中，女性携带者表现出正常和异常组织的混杂状态。偶见的**显性杂合子**（manifesting heterozygote）被视为 X 连锁隐性疾病。这些女性可能受累非常严重，因为一些重要组织中大部分细胞已经非常不幸地失活了正常那条 X 染色体。

Y 染色体携带相对少的基因

除了男性自身外，无已知的人类性状会产生图 4.5E 所示的常见的 Y 连锁系谱（对“豪猪男”和“多毛耳”的声明是不可信的，分别见 MIM 146600 和 425500）。由于正常女性缺乏所有的 Y 连锁基因，因此，任何这些基因一定编码不重要的性状或是决定男性特异的功能。有些基因在 X 和 Y 染色体上以功能的拷贝形式存在；他们对这种争论可能证明了一个例外，但是不会产生经典的 Y 连锁系谱方式。Y 染色体长臂的微缺失是男性不育的一个重要原因，但是，不育的男性不会表现出图 4.2E 那样的家系。Jobling 和 Tyler-Smith（2000）及 Skaletsky 等（2003）总结了 Y 染色体的基因成分和其在疾病中的可能涉及的情况。

位于 Xp-Yp 配对区的基因显示假常染色体遗传

如节 2.3.3 中所述，Xp、Yp 远侧的 2.6Mb 是同源的，并在减数分裂时易于发生交叉。故这些区域的基因单独表现为**“假常染色体”**（pseudoautosomal）而不是性连锁遗传方式。

4.2.3　在单个系谱中很难明确地确定遗传方式

假如一个人类家系大小有限，单纯凭借观察一个系谱图来完全确定某一性状的遗传方式不大可能。人们用实验动物进行杂交试验，验证比例为 1∶2 还是 1∶4。对于人类系谱，受累患儿的比例并不是十分可靠的指标。这主要是因为家系成员的数目太少，此外，是因为确定家族的方法可能偏移了患者与正常后代的比例。对于隐性疾病来说，患儿的比例通常看似大于 1∶4。这是因为当家系中有一个受累患儿该家系一般就被确认了；而双亲都是携带者，但所幸无受累者的家系则被系统性漏检。这些**确证性偏移**（biase of ascertainment）及其纠正方法在节 15.2.1 中会讨论。

对于许多更为少见的疾病，所描述的遗传方式仅仅是一个告知性的猜测。确定遗传方式很重要，因为它是用于遗传咨询中风险评估的基础。然而重要的是，要认识到遗传方式通常只是有效的假说而不是既定的事实。OMIM 用星号表示相对建立完善的遗传

方式条目。只有在得到该基因的克隆拷贝之时才能确定其到底是哪一种遗传。

4.2.4　一个基因一个酶并非是一个基因一个综合征

系谱方式给人类遗传学提供了必要的切入点，但对于确定基因来说它们仅是一个起点。认为约 6000 多个已知的孟德尔性状就确定 6000 个 DNA 编码序列是个严重的错误。这是对 Beadle 和 Tatum 的**一个基因一个酶假说**（one gene-one enzyme hypothesis）的无根据扩展。追溯到 20 世纪 40 年代，这个假说让人们在理解基因如何决定表型上有了大步的跨越。从那时起该假说就被扩展：有些基因编码非翻译的 RNA，有些蛋白质并不是酶，很多蛋白质含有几个独立编码的多肽链。但即使有这样的扩展，Beadle 和 Tatum 的假说也不能提示 OMIM 目录中的条目和 DNA 转录单位之间有一一对应的关系。

经典遗传学的基因是抽象的东西。任何由单个染色体位置所决定的性状都会以孟德尔方式分离出来——但是，此决定因素可能不是分子遗传学家所说意义上的基因。Fascio-scapulo-humeral 肌营养不良（某些肌群存在严重的但非致死性肌无力；MIM 158900）是 4q35 上序列的小缺失引起，尽管已在此区域进行了集中搜索和测序，但在本书撰写之时，尚无人在该区域找到相关蛋白的编码序列。Charcot-Marie-Tooth 病 1A 型（运动和感觉神经元病；MIM 118220）的致病“基因”经证实是染色体 17p11.2 上的 1.5Mb 片段的串联重复（节 16.5.2）。这些例子并不常见——多数 OMIM 条目可能只是描述影响单个转录单位的突变序列。然而，由于三种异质性，表型和转录单位之间仍不是一一对应关系。

- ▶ **基因座异质性**（locus heterogeneity）　是指相同的临床表型可由不同基因座上的突变引起。
- ▶ **等位基因异质性**（allelic heterogeneity）　患有特定遗传病的不同患者可见在一既定的基因内发生许多不同的突变（在 16 章更充分地讨论）。很多疾病既表现为基因座异质性也表现为等位基因异质性。
- ▶ **临床异质性**（clinical heterogeneity）此处用于描述同一基因突变产生的两种甚至更多种不同的疾病的情况。节 16.6 举例说明。

基因座异质性常见于复杂通路破坏所致的综合征中

听力丧失提供了基因座异质性的好例子。当两个常染色体隐性的深度先天性听力丧失患者婚配，正如他们通常会这样，他们的子女大多具有正常的听力（图 4.3）。显而易见，要构成一个耳蜗毛细胞一样精密的“机器”需要很多不同的基因，任何一个此类基因的损伤都能导致听力丧失。无论何时父母的不同基因携带突变，子女都具有正常的听力。这是**互补作用**（complementation）的一个例子（框 4.2）。这种基因座也异质性只在听力丧失、失明、智力低下这类相当普遍的通路破坏的疾病中所料见；但是即使伴有更特殊的病理原因，多个基因座也是很常见的。一个显著的例子是 Usher 综合征，合并听力丧失、渐进性失明（色素异常性视网膜炎）的常染色体隐性疾病；它可能由 10 个或更多个不连锁的基因座突变引起（遗传性听力丧失网页：www.uia.ac.be/dnalab/hhh/）。OMIM 就已知的基因座异质性的例子有单独记录（通过连锁或突变分析来

确定)，但是一定还有许多包含在单独条目中未被发现的例子。

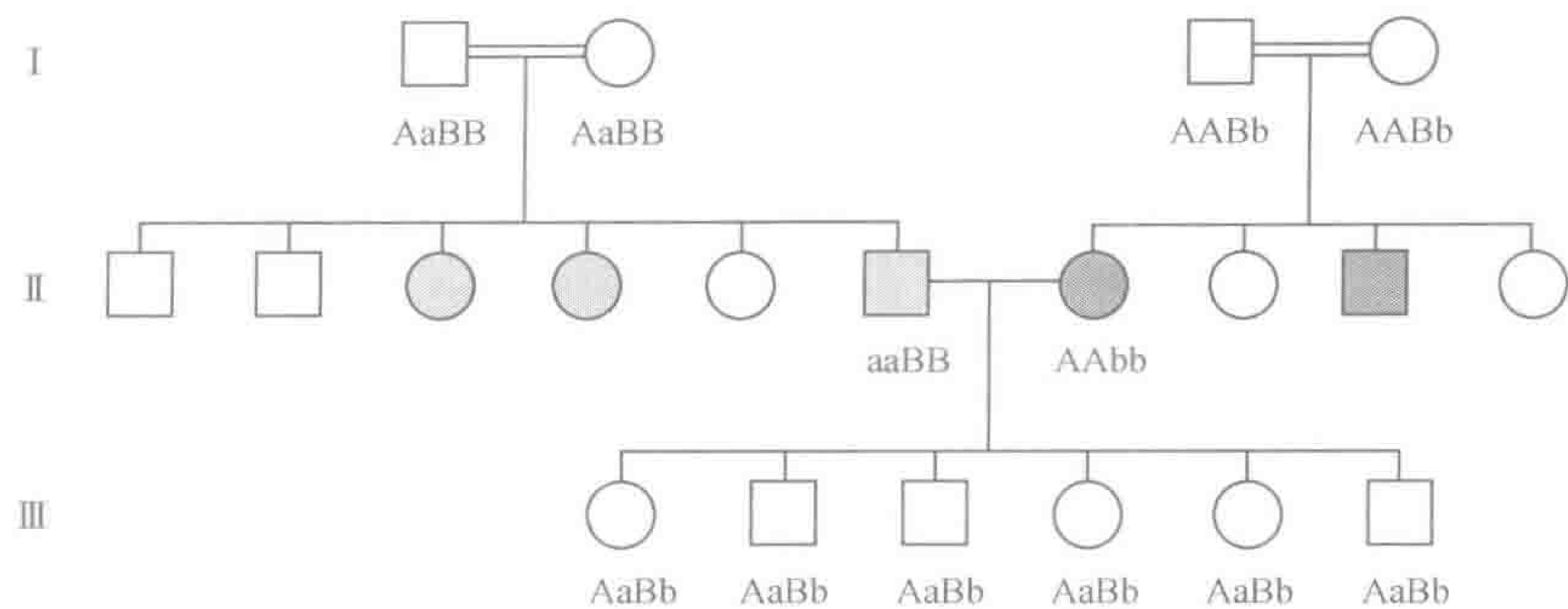


图 4.3　互补作用：常染色体隐性深度听力丧失的双亲通常生出听力正常的孩子 $Ⅱ_6$ 和 $Ⅱ_7$ 是未受累但为近亲婚配双亲的孩子，二者均有受累同胞，使得每个后代都有可能患常染色体隐性听力丧失。他们的孩子均未受累，说明 $Ⅱ_6$ 和 $Ⅱ_7$ 有非等位基因突变。

框 4.2　互补实验可发现两种隐性性状是否由等位基因决定

	一个基因座	两个基因座
双亲杂交	$a_1a_1 \times a_2a_2$	aaBB×AAbb
子代	a_1a_2	AaBb
表型	突变	野生型

两种性状的纯合子动物杂交，并观察子代的表型。如果两种动物在同一基因座携带突变，其后代将没有野生型的等位基因，因此表型异常。如果有两个不同基因座存在突变，后代是这两种每一个隐性性状的杂合子，所以表型是正常的。在同一基因座上极少有等位基因彼此互补（等位基因间的互补作用）。

等位基因系列是引起临床异质性的原因

有时一些有显著区别的人类表型已证明是由同一基因座上不同等位基因突变引起的。其差别可能是程度上的差别：如，抗肌萎缩蛋白基因部分失活的突变导致 Becker 肌营养不良，而同一基因完全失活的突变则导致类似但病情更为严重的 Duchenne 肌营养不良（致死性肌肉消瘦：MIM 310200）。其他情况下这种差别是质的差别：雄激素受体基因失活引起雄激素不敏感（核型为 46，XY 的胚胎发育为女性；MIM 313700），但是在同一基因的谷氨酸密码子扩张则引起一种显著不同的疾病，脊髓延髓肌营养不良或 Kennedy 病（MIM 313200）。这些及其他有关于基因型和表型的相互关系将会在 16 章深入探讨。

4.2.5　线粒体遗传表示为可识别的母系系谱方式

除细胞核染色体上的基因突变外，线粒体突变也是引起人类遗传疾病的重要原因（详细内容见线粒体图数据库：http：www.mitmap.org）。线粒体基因组（节 9.1.2）

虽然很小但与核 DNA 相比有很高的可突变性，可能是因为 mtDNA 复制更容易出错、复制次数更多。线粒体类疾病有两个不寻常的特点：**母系遗传**（matrilineal inheritance）和常见的**杂质性**（heteroplasmy）。

因为精子并不为合子提供线粒体，故线粒体遗传是母系遗传（这个结论有赖于有限的证据；但是在孩子中几乎从未检测到父源线粒体变异体）。因此既定的一个可识别的系谱方式，线粒体遗传疾病影响两性，但只通过受累母亲传递给后代（图 4.4）。

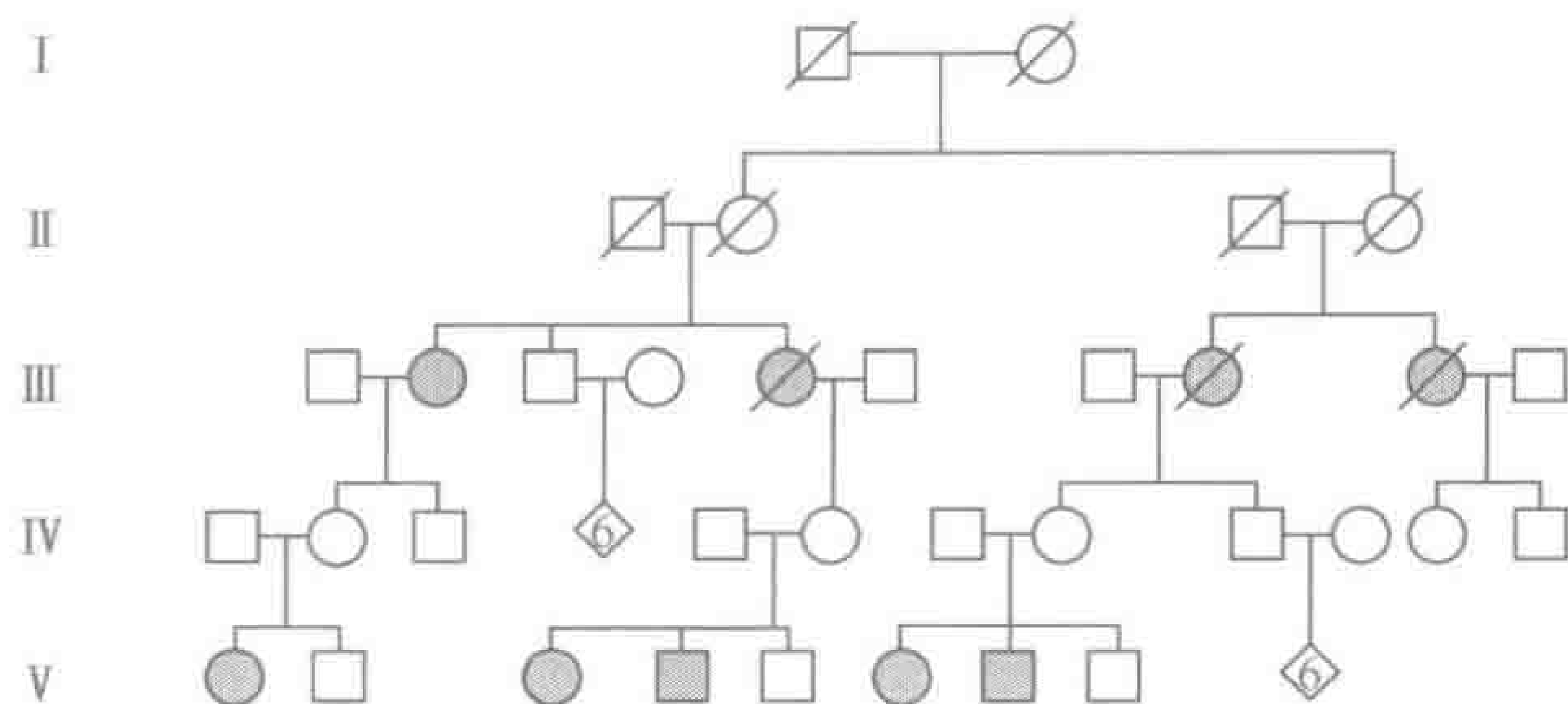


图 4.4 线粒体疾病的系谱

该典型系谱图显示一个由线粒体异常所致的听力丧失（该家系由 Prezant 等 1993 年报道）。注意不完全外显。

细胞含有多个线粒体基因组。在有些线粒体病的患者，每个线粒体基因组都携带致病突变（**纯质性**，homoplasmy），但在其他病例中，在每个细胞可见正常的和突变的线粒体基因组的混合群（异质性）。不像核遗传嵌合体必将引起合子后的嵌合体（节 4.3.6），线粒体杂质性可以从杂质性母亲传递给杂质性后代。在这样的病例中，在母亲和子代之间，异常线粒体基因组的比例可有显著不同，这表明相当少量的母亲的线粒体 DNA 分子产生后代所有线粒体 DNA（节 11.4.2）。线粒体疾病的复杂分子病理学将在节 16.6.6 中讨论。

4.3 基本的孟德尔式系谱方式中的复杂情况

在现实生活中，各种复杂情况通常看似为基本孟德尔遗传方式，图 4.5 显示一些常见的复杂情况。

4.3.1 常见隐性疾病可能显示为假显性系谱方式

若某一性状在人群中常见，则有很高的几率由两个或更多的人独立地组成该系谱。一种常见隐性性状如 O 血型会因 O 型血的人与杂合子多次的婚配而见于连续数代成员。这就产生了类似显性遗传方式（图 4.5A），因此经典的系谱方式最好见于罕见的性状中。

4.3.2 显性性状未表现出来称为“不外显”

对显性疾病来说，**不外显**（nonpenetrance）是一种常见的复杂情况。对于给定的基因型，性状**外显率**（penetrance）定义为这种基因型的人表现为该性状的可能性。从

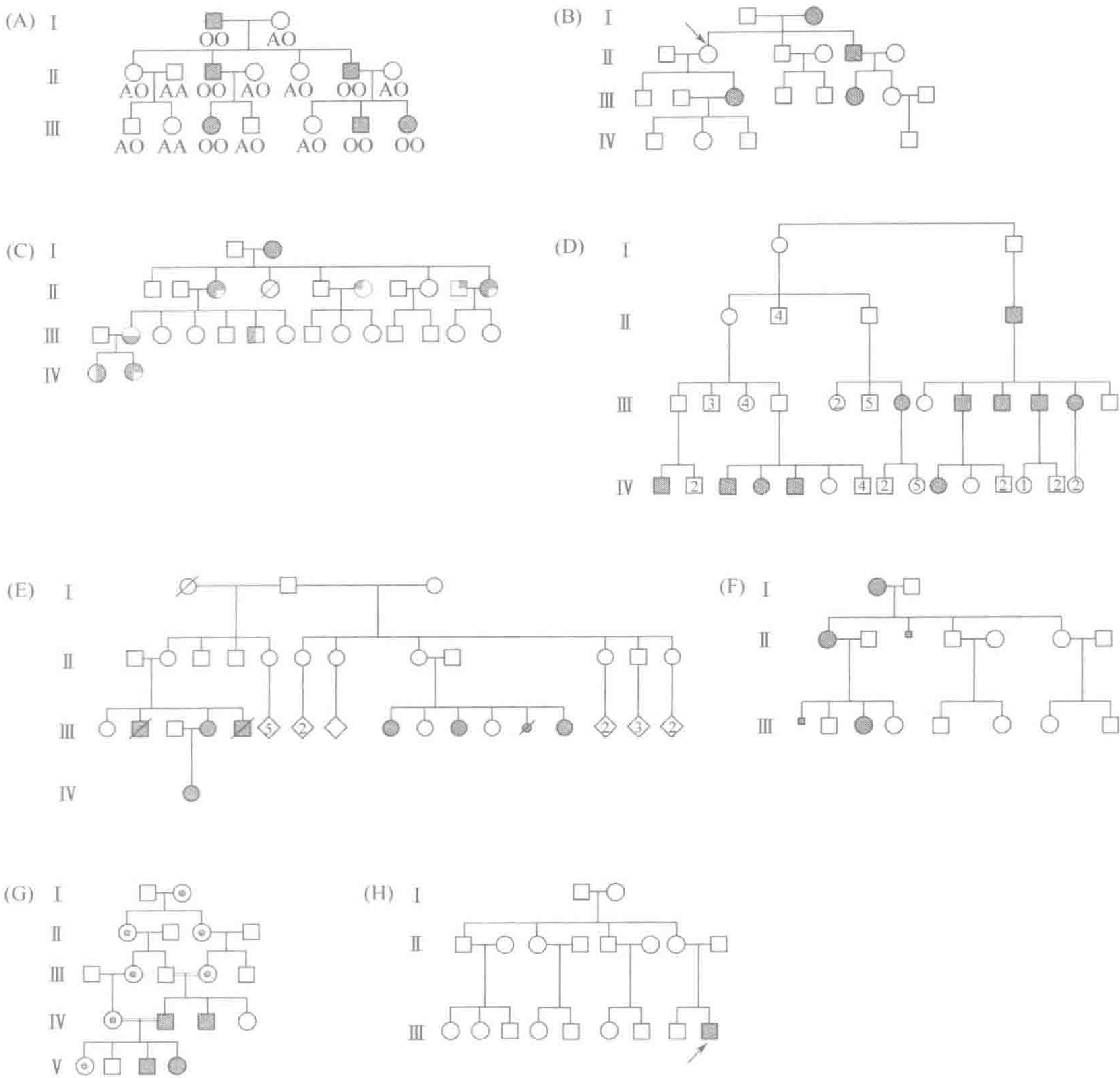


图 4.5　基本孟德尔遗传方式中的复杂情况

(A) 常见隐性诸如血型 O 可能表现出显性方式。(B) 常染色体显性遗传中 II_2 不外显。(C) 常染色体显性遗传中的不规则表达：在此 Waardenburg 综合征家系中，第一个象限阴影＝耳聋；第二个象限阴影＝眼睛颜色不同；第三个象限阴影＝前额白发；第四个象限阴影＝头发早白。(D) 遗传印记：该家系中，常染色体显性遗传的血管瘤只有在该基因由父亲遗传而来时才表现为该病（家系由 Heutink 等报道，1992）。(E) 遗传印记：常染色体显性遗传的 Beckwith-Wiedemann 综合征，只有在该基因由母亲遗传而来时才表现为该病（家系由 Viljoen and Ramesar 报道，1992）。(F) X-连锁显性的色素失调症。受累男性胎儿自发流产（以小方形表示）。(G) X-连锁隐性的系谱中，近亲婚配后代有一位受累女性，且表现出明显的男性-男性传递的现象。(H) 一种新的常染色体显性突变，类似常染色体隐性或 X-连锁隐性遗传。

定义看来，显性性状在杂合子的人表现出来，因此，应该显示为 100％的外显率。无论如何，通常显示为显性遗传的许多人类性状，偶尔也会跨过一代。在图 4.5B 中，II_2有受累的父母和孩子，并几乎是肯定携带突变基因，但是表型上却正常。这将描述为不外显的例子。

不外显并不稀奇——的确，100%外显才是更令人惊讶的现象。在总的和正常的环境下，性状的表现与否经常取决于该基因座上的基因型，但若是异常的遗传背景、特殊的生活方式或可能恰恰出于偶然，意味着时常有人不表现出该性状。不外显是遗传咨询中的主要陷阱。已知图 4.5B 的疾病为显性、III_7无症状而告诉她没有生出患儿的风险，这个咨询师是不明智的。遗传咨询师的工作之一是了解每种显性症状的通常外显度。

当然，某性状经常依赖于多个因素，即使完全是遗传性的也不显示为孟德尔系谱方式。随着其他遗传基因座和/或环境的影响增加，从完全外显的孟德尔遗传到多因子遗传有一个性状连续性（图 4.6）。无逻辑上的分界把不完全外显的孟德尔性状与多因子性状分离开来：这是一个对应用来说最有用的描述。

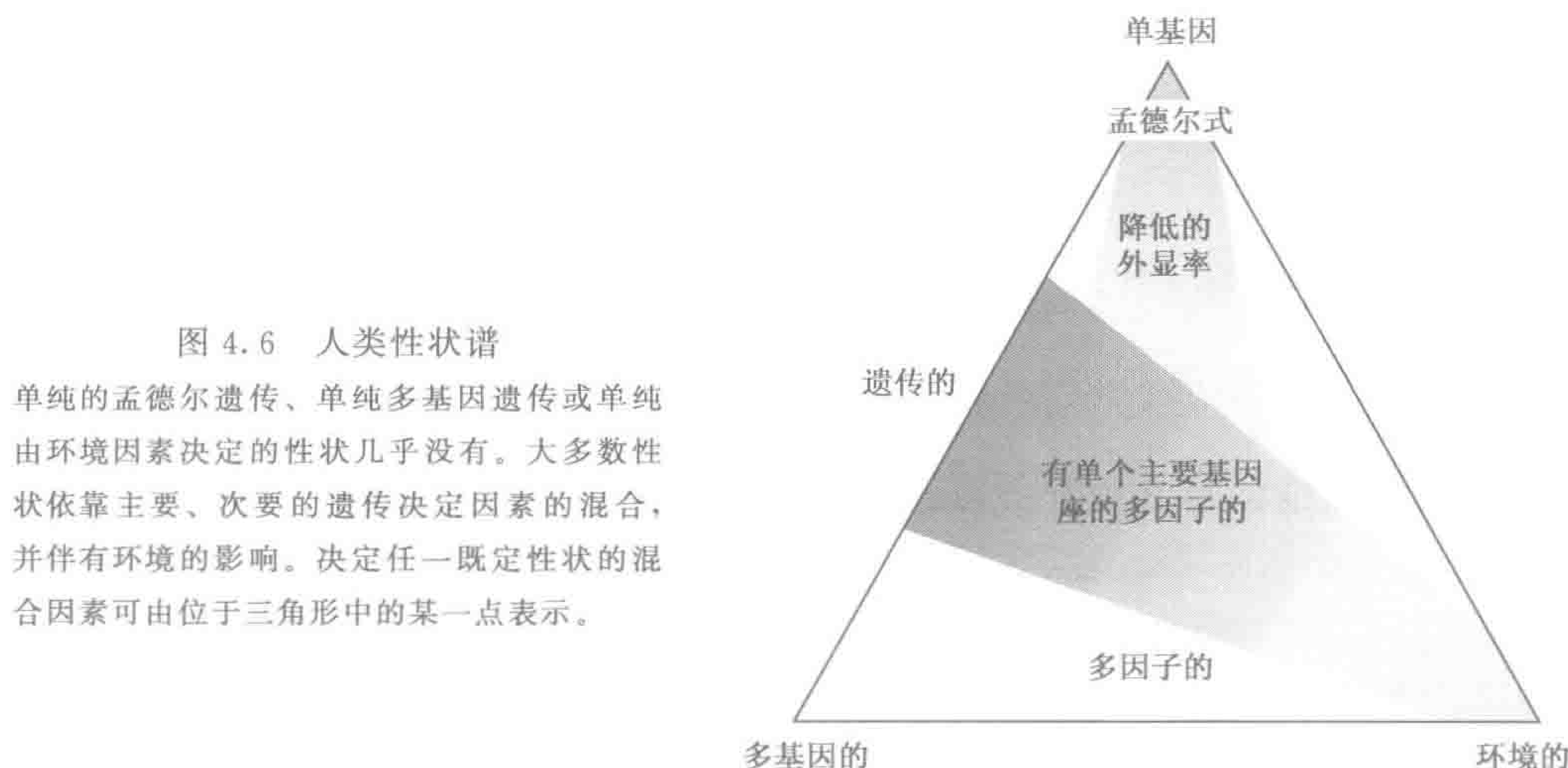


图 4.6　人类性状谱
单纯的孟德尔遗传、单纯多基因遗传或单纯由环境因素决定的性状几乎没有。大多数性状依靠主要、次要的遗传决定因素的混合，并伴有环境的影响。决定任一既定性状的混合因素可由位于三角形中的某一点表示。

迟发病中与年龄相关的外显率

特别重要的外显率降低的例子见于迟发性疾病。遗传病并非必是先天性的（出生即表现出来）。妊娠时基因型是固定的，但表型可能直到成年才表现出来。在这类病例中外显率与年龄相关。亨廷顿（Huntington）病（进行性神经退行性变：MIM 143100）是个典型的例子（图 4.7）。延迟发病可能是由毒性物质缓慢积累引起的，由慢性组织坏死或不能修复的一些环境损伤造成的。遗传性癌是由人的细胞中已有一个突变的肿瘤抑制基因又发生第二次突变所引起的（17 章）。根据疾病而言，只要此人活得足够久，其外显率就变成 100%，或者也有些携带致病基因的人无论活多久也不产生症状。像图 4.7 发病年龄曲线是遗传咨询中的重要工具，因为它帮助遗传学家预计有发病风险但无症状的人最终将有多大几率发病。

4.3.3　许多疾病的表现度不一致

与不外显相关的**表现度不一致**（variable expression）常见于显性性状。图 4.5C 列举了一个 waardenburg 综合征家系的例子。家系不同成员表现出疾病的不同特点。其原因是不外显的原因：其他的基因，环境因素，或纯粹的偶然因素对症状发生有一些影

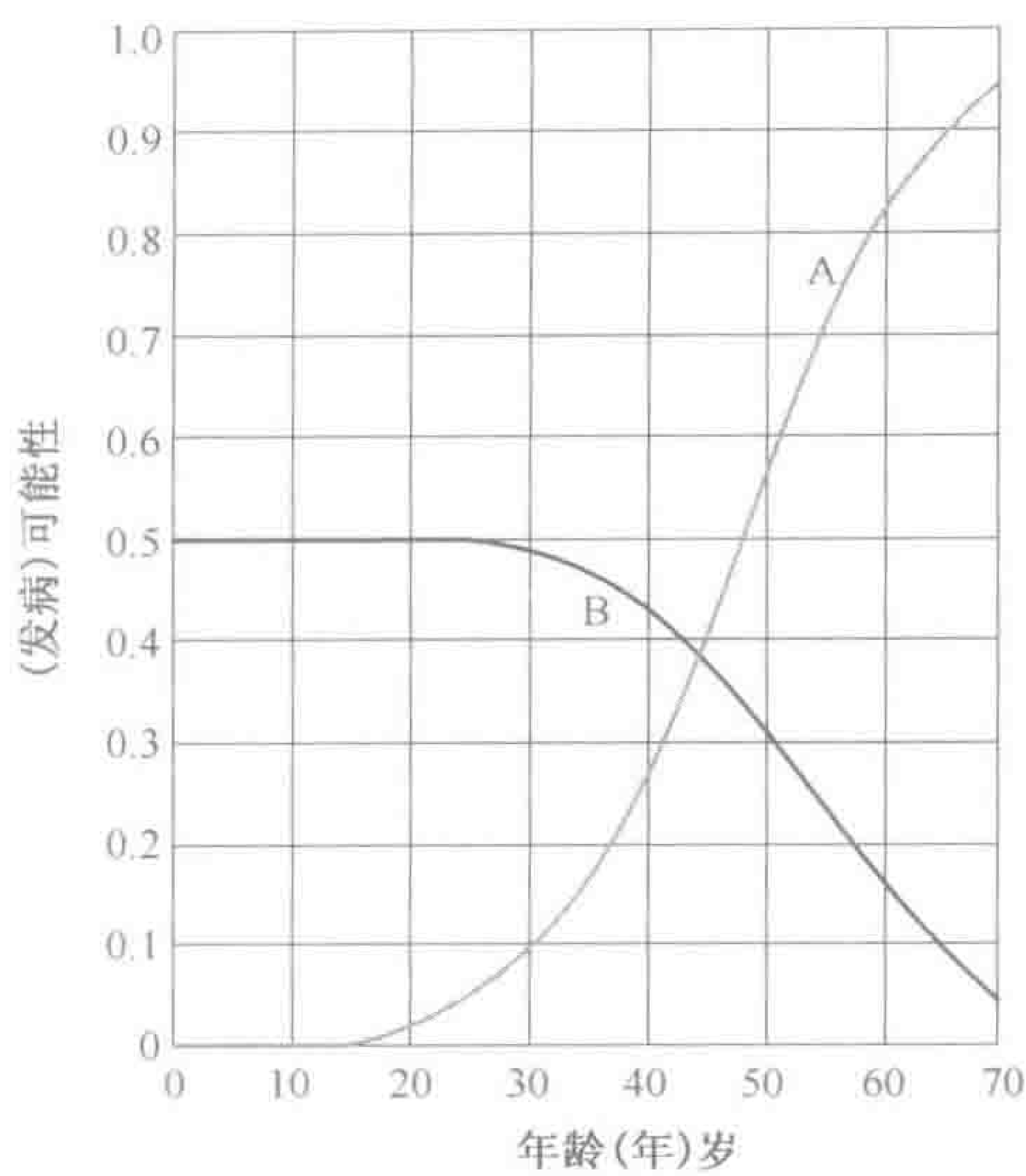


图 4.7　Huntington 病发病年龄曲线
曲线 A：带有致病基因的个体到一定年龄发病的可能性。曲线 B：受累双亲的一个健康的孩子在一定年龄携带致病基因的风险。引自 Harper (2001)。Practical Genetic Counselling，第五版，Hodder Arnold。引用经 Hodder Arnold 许可。

响。不外显和表现度不一致是显性而非隐性性状的典型问题。这部分反映了在典型的隐性系谱中确定不外显的困难所在。然而总体来说，隐性疾病没有显性的变异多，可能因为杂合子表型涉及两个等位基因效果的平衡，故其对外界影响比纯合子表型更敏感。然而，不外显与表现度不一致时二者偶尔也可见隐性疾病。

这些复杂情况在人身上没有在植物和其他的动物表现明显。实验动物和农作物比人的遗传一致性更强。我们在人类遗传学中所观察的是典型的野生群体。无论如何，鼠遗传学家熟悉个体处于不同遗传背景时突变表达发生变化的方式——这正是在研究人类疾病的小鼠模型时要重点考虑的。

早现是表现度不一致的特殊类型

早现（anticipation）描述的是有些可变的显性疾病在世代传递中变得更加严重（或更早发病）的趋势。直到最近，大多数遗传学家都怀疑这种情况是否真正发生过。问题是真正的早现非常容易被严重病例中的随机变异所迷惑。当家族有病情严重的患儿出生时，就要引起注意了。在调查病史时，遗传学家注意到患者双亲之一受累，但病情较轻微。这种情况看起来像是早现，但实际上可能恰恰就是确证性偏移。要是父母已严重受累，那么他或她可能将永远不能做父母了；要是子女病情轻，该家系就不会引起注意。假设缺少解释早现和面临这些偏移显示为早现的统计学问题的合理的机制，大多数遗传学家不愿意把早现看得那么严重——直到早现的分子研究发展起来，才迫使他们重视这个问题。

随着脆性 X 综合征（具有不同体征的智力低下；MIM 309550）的三核苷酸不稳定扩张及后来强直性肌营养不良（变化多端的系统性疾病，特征性肌肉营养不良；MIM 160900）和亨廷顿病的发现，早现突然变得备受瞩目甚至成为研究的热点。这些疾病的严重程度和发病年龄与核苷酸重复片段长短相关，重复片段长度随着基因世代传递变长（节 16.6.4）。这些疾病显示为真正的早现。如今我们再次回顾造成许多疾病的早现，重要的是切记，确证性偏移这个陈旧异议仍是有根据的。早现的确定不仅依靠临床表现

还需要详细的统计学支持才可信。

4.3.4　印记基因的表达取决于基因的亲代来源

某些人类性状是常染色体显性遗传，两性均可受累，而且基因可由父母任一方传递给后代，但是只有特定性别的某一方遗传才会表现出来。例如（图 4.5D）在常染色体显性遗传的胶质瘤家系中，只有该基因是由父亲遗传给子女才发病，而 Beckwith－Wiedemann 综合征（先天脐疝，巨舌，过度生长；MIM 130650）有时是显性的，但只有在致病基因由母亲遗传而来的子女才发病（图 4.5E）。这些父母性别的影响就是**印记**(imprinting) 的证据，难以理解的现象是特定基因是如何被标记（印记）上其双亲的来源。围绕印记的机制及进化目的等众多问题在节 10.5.4 中讨论。框 16.6 中列举了临床上典型的病例。

4.3.5　男性致死性使 X 连锁的系谱复杂化

对于一些 X 连锁显性遗传病，缺少正常等位基因在出生前是致死的。因此受累男性没有出生，而我们看到的是只有女性受累的疾病，且她们把疾病传递给一半的女儿，但不传递给儿子（图 4.5F）。可能有流产史，但其家系很少大到足以证明男孩的数目仅仅是女孩的一半。一个例子是色素异常（沿着特定类型的被称为 Blaschko’s 线的线性皮肤缺损，常伴神经或骨骼异常；MIM 308310）。

4.3.6　新的突变经常使系谱的解释复杂化，并能产生嵌合体

许多严重的显性或 X-连锁遗传病的病例是新突变的结果，这使那些先前无此种病史的家系在毫无预兆的情况下感到震惊。完全外显的致死显性将总是由新突变而发生，而父母从未受累——致死性发育不良就是个例子（长骨严重变短，颅缝异常融合；MIM 187600）。对于非致死性但严重的显性疾病来说，一个同样的理由但仅在严重程度低一些的情况适用。如果疾病使大多数受累者无法生育，但该病的新病例仍然持续发生的话，那么大多数新病例一定是由新突变引起的。严重的 X 连锁隐性也显示出大部分的新突变，因为在男性任何时候其基因都要面临自然选择。另一方面常染色体隐性系谱未受明显影响——一个突变等位基因可以在无症状携带者中传递多代，我们可毫无异议的断定受累子女的父母都是携带者。

假定平均一段时间，新的突变恰恰替代了通过自然选择丢失的疾病基因，人群中替代的频率有赖于自然选择淘汰有害基因的速度与新发突变产生有害基因的速度之间的简单关系（节 4.5.2）。影响群体等位基因频率的一般机制在第 11 章框 11.2 中讨论。

当无相关家族史的健康夫妇生出严重异常的后代时（图 4.5H），确定遗传方式和再发风险是非常困难的：问题可能是有一新突变的常染色体隐性、常染色体显性，X-连锁隐性（如果后代是男性），或是非遗传性的。更复杂的情况是由生殖嵌合性引起的（见下文）。

嵌合体有两个（或多个）遗传上不同的细胞系

我们已经发现，在严重的常染色体显性及 X-连锁疾病中，受累者子女很少或没有

子女，在人群中致病基因由再发突变来维持。通常的假说是一个完全正常的人产生一个单突变的配子。然而，这并非是必定发生的事。除非突变过程有些特殊，诸如突变只在配子形成期间发生，突变可以发生在合子后的生命的任何时候。合子后的突变产生有两个（或多个）遗传上有差异的细胞系的**嵌合体**（mosaics）。

嵌合性能影响体细胞和（或）生殖细胞系组织。合子后突变不仅频发，而且是不可避免的。人的突变率在每个基因、每个细胞世代一般为 10^{-7}，而我们机体含有大约 10^{13}个细胞。由此得出对于不计其数的遗传病来说，我们每个人一定都是嵌合体。的确，正如 John Edwords 教授经典的评价那样，一个健康男人每次射精会产生全部 OMIM 目录的遗传病。这种说法不应该引起惊奇。如果你手指上的一个细胞突变为亨廷顿病的基因型，或者耳朵上的一个细胞携带囊性纤维化突变，对你或你的家庭根本没有影响。只有体细胞突变导致突变细胞真正的克隆的出现，整个机体才有风险。这可能以两种方式发生：

- 突变导致原本复制慢或根本不复制的细胞异常增生，因而产生了突变细胞的克隆。这是癌发生的方式，这一整个论题将在 17 章详细讨论。
- 突变发生于早期胚胎，累及的细胞是整个有机体中重要分裂的祖代。那种情况下嵌合性个体可能表现出疾病的临床症状。

产生于双亲的生殖系突变可能导致始于其子女的遗传疾病。早期生殖系突变能产生携带大量突变的生殖系细胞克隆（**胚胎的-或性腺的-嵌合体**）（germinal-or gonadal-mosaicism）的个体。结果，先前无家族史的健康夫妇可能生出不止一个患有同样严重的显性疾病的后代。系谱看似隐性遗传。即使已判断出准确的遗传方式，计算再发风险用于双亲咨询也是相当困难的（van der Meulen *et*.，1995）。通常要引用经验风险值（节 4.4.4）。图 4.8 举例说明 X 连锁疾病的病例中嵌合性导致的遗传咨询的不确定性。

分子水平的研究对这些病例帮助很大。有时可以直接证明健康父亲产生相当比例的

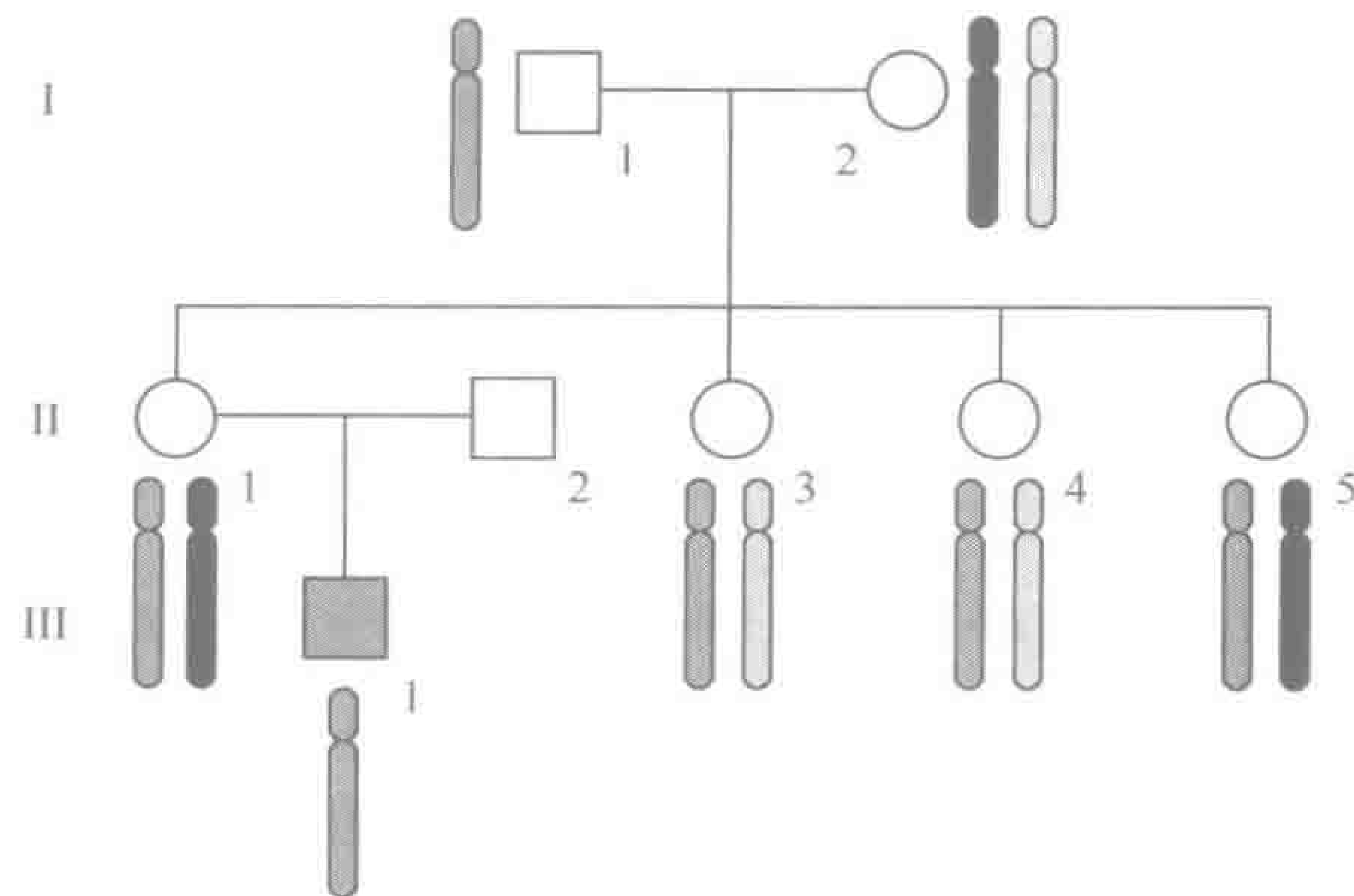


图 4.8　X-连锁隐性遗传的杜兴肌营养不良中的新突变

三个祖父母源的 X 染色体已用遗传标记区别开，以深灰、浅灰、黑色显示（忽略重组）。Ⅲ₁ 具有祖父源的 X 染色体，在系谱中该染色体在某些点获得一个突变。有四个可能的点已发生突变：如果Ⅲ₁ 携带一个新突变，整个家系成员的再发病风险很低；如果Ⅱ₁ 是生殖系嵌合体，其未来子女有很高的发病风险（但是难以量化），但她的姐妹们不会；如果Ⅱ₁ 是单个突变精子的结果，对 X 连锁隐性性状她有标准的再发风险，但其姐妹无再发风险；如果Ⅰ₁ 是生殖系嵌合体，所有的姐妹有难以量化的很大发病风险。

突变精子。直接检查女性的生殖系细胞是不可能的，但检测如成纤维细胞或发根这些易获得的组织也能得到嵌合体的证据。身体组织检查结果阴性并不排除生殖系嵌合体，但一阳性结果且有一受累孩子则证实为生殖系嵌合体（图 4.9）。

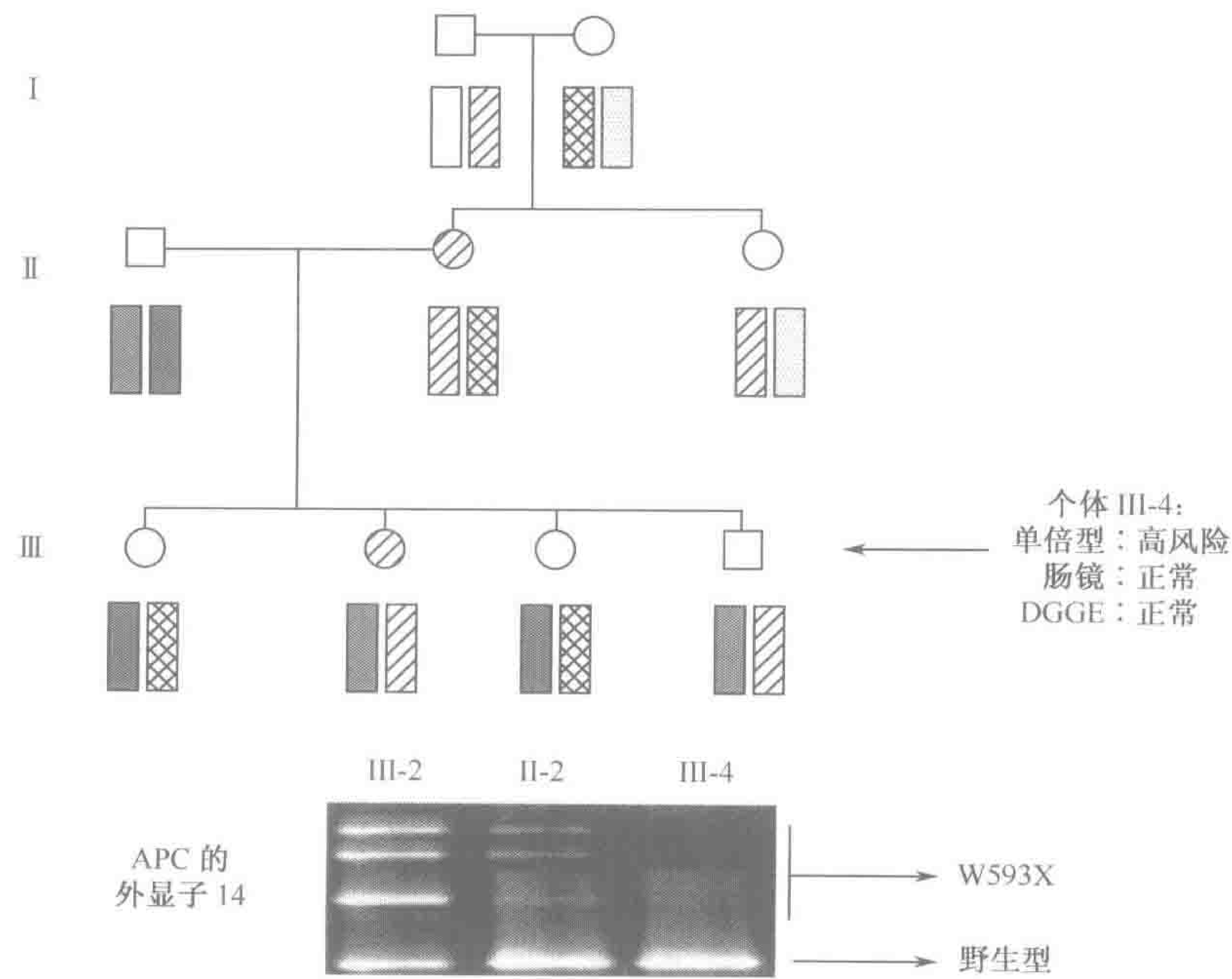


图 4.9　显性遗传病中的生殖系和体细胞嵌合体

Ⅱ$_2$ 和Ⅲ$_2$ 患有家族性腺瘤息肉病，为显性遗传性结直肠癌，定位于 5 号染色体（MIM 175100，节 17.5.3）。Ⅱ$_2$ 双亲不受累。变性梯度凝胶电泳（节 18.3.2）显示突变 W593X 在Ⅲ$_2$ 的 APC 基因外显子 14 上（凝胶 1 泳道上方的条带）。对Ⅱ$_2$ 凝胶显示突变带，但非常微弱，说明此人的血样是突变的嵌合体。在Ⅲ$_4$ 无突变，甚至通过研究其连锁标志（编码结肠的染色体，节 18.5）显示，他遗传了其母亲的高危染色体（ ）。临床检查（结肠镜）证实Ⅲ$_4$ 不患该病。Ⅱ$_2$ 一定是生殖系和体细胞突变的嵌合体。病例和凝胶由 Bert Bakker，Leiden 教授惠赠。

异源嵌合体有单一机体的两个单独合子的细胞

嵌合体（mosaic）的生命始于一个受精卵。另一方面，**异源嵌合体**（chimera）是两个合子融合形成一个胚胎的结果（与双生子相反），或来自于非同卵双生子细胞的一个双生子选择性地有限克隆化（图 4.10）。异源嵌合性通过存在几个基因座上双亲太多的等位基因共享的组织标本所证实（若只涉及一个基因座，则推测是单个突变的嵌合体）。血型中心偶尔会在健康献血者中发现嵌合体，有的两性畸形患者经证实是XX/XY嵌合体。例如，strain 等（1998）叙述了一个 46，XY/46，XX 的男孩是体外受精的三个胚胎植入母体后两个胚胎融合的结果。

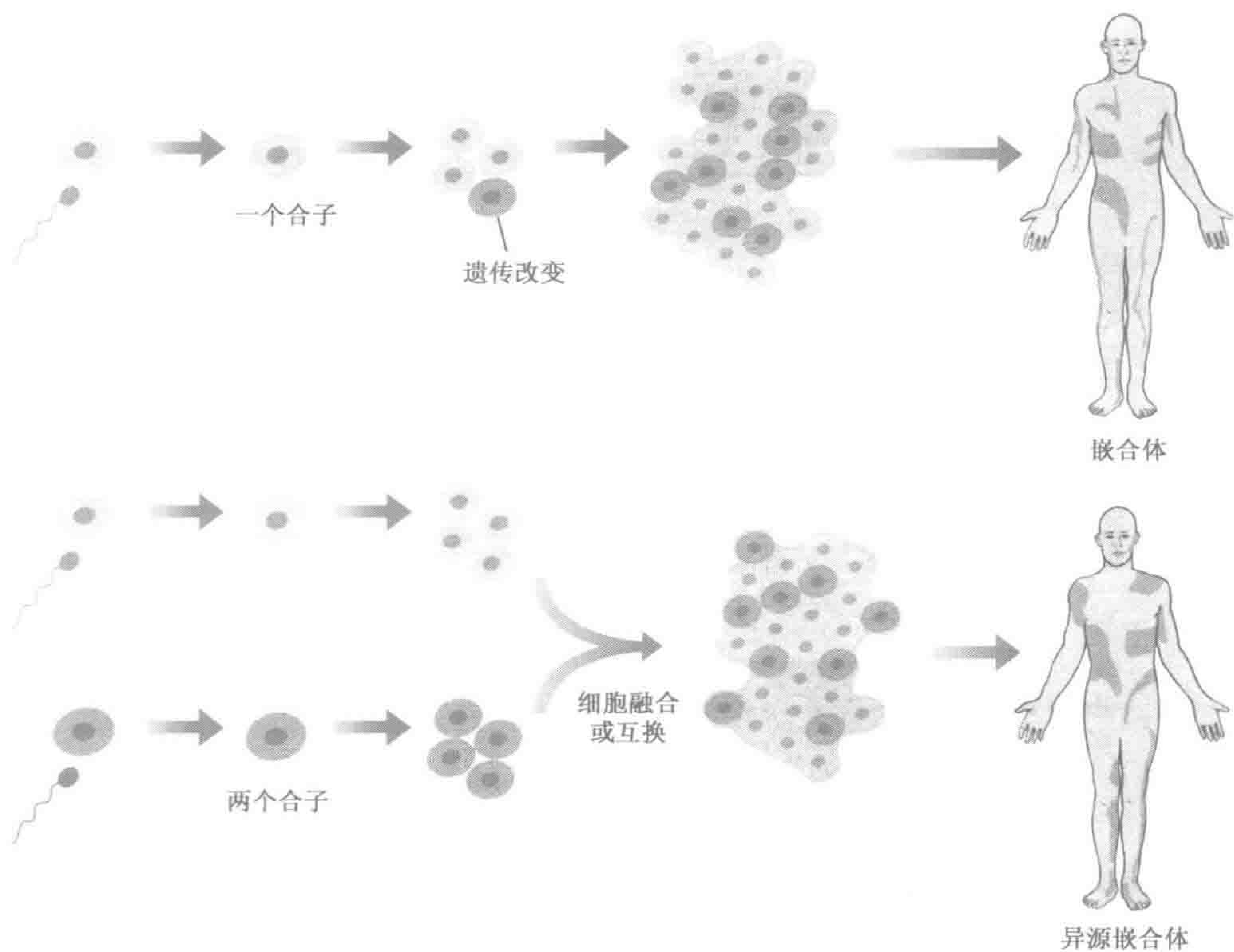


图 4.10　嵌合体与异源嵌合体

嵌合体具有源于单个合子的两个或更多遗传上不同的细胞系。其遗传改变可能是基因突变，染色体数目或结构的改变、或者是莱昂作用的特殊病例中 X 染色体失活。异源嵌合体来源于通常正常但遗传上有差别的两个合子。

4.4　多因子性状的遗传学：多基因的阈值理论

4.4.1　历史回顾

当孟德尔的研究于 20 世纪重新被发现时，英国和其他各地已建立了相竞争的遗传学学院。Francis Galton 是 Charles Darwin 出名的古怪堂弟，他把他的天赋都用在系统化研究人类的变异现象上。首先关于“遗传的天资和性状”（Hereditary Talent and Character）的论文和孟德尔的论文于 1865 年同年发表（1869 年扩编为专著 Heredity Genius），他花费多年时间调查家系成员的相似性。Galton 潜心于量化观察结果和运用统计学分析。1884 年于伦敦建成了他的人类统计实验室，记载了他的研究对象（给予他三便士的酬劳）的体重、坐高、站高、臂长、呼吸容量、肌肉推挤力量、击打力、反应时间、视力和听力的灵敏性、颜色分辨力和距离判断力。首次应用统计学之一，是他比较了父母和子女的自然特征，确立亲属之间的亲缘度。至 1900 年，他已经积累了大量的性状的遗传资料，建立了他们调查研究的传统方法（**生物统计学**，biometrics）。

当再度发现孟德尔的研究工作时引发了争议。生物统计学家承认，孟德尔式基因能解释一些少见的异常或奇特的怪癖。但是他们指出，大多数性状（身高、体质、肌力、

发现猎物或寻找食物的本领）可能在进化中是重要的，是连续的或数量的性状，经不起孟德尔遗传分析的检验。我们都具备这些性状，只是程度上有差异，因此不能用画系谱和标明哪些人有此性状来确定遗传方式。孟德尔式分析需要你或有或无的**“双歧性状”**（dichotomous character）。孟德尔学家与生物统计学家之间的争议一直持续到 1918 年并不时激化。同年，Fisher 的学术论文表明，正如生物统计学家所描述的那样，由大量独立的孟德尔因子［**多基因**（polygenic）性状］所控制的性状将表现为精确的连续特性、数量的变异和家族相关性。后来 Falconer 将模型扩展为涵盖双歧性状。Fisher 和 Falconer 的分析创造了人类遗传学上统一的理论基础。下面的各节用非数学的形式陈述他们的观点。更有力的论述，例如 Falconer 和 Mackay 的，在 1996 年有关数量遗传学的教材中都能够找到（进一步阅读）。

4.4.2　数量性状的多基因理论

任何依赖于众多独立微小因素的加性效应的可变性状在人群中将显示为**正态（高斯）分布**［Normal（Gaussian）distribution］。图 4.11 给予这一正态分布非常简化的说明。我们假定某性状依赖于一个、两个或三个基因座的等位基因。当更多的基因座包括在内，我们就会发现两个结果：

▶ 基因型和表型之间简单的对应关系不复存在了。除非有极端表型，否则不可能从表

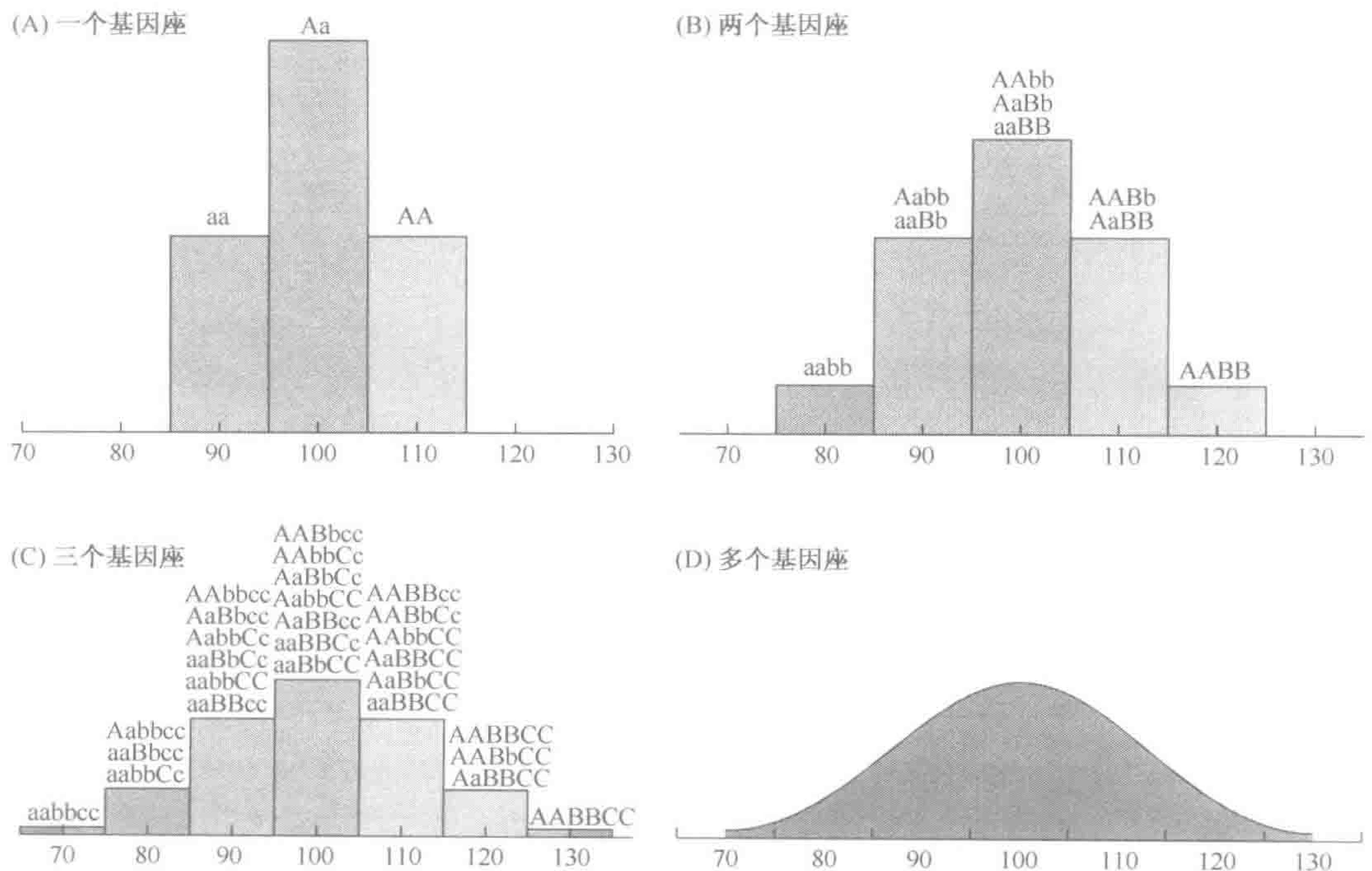


图 4.11　逐渐的近似高斯分布

此图表显示某均值为 100 单位时，一个假设性状在人群中的分布。该性状由等位基因的加性效应（共显性）决定。每个上调等位基因使该值增加 5 个单位，每个下调的等位基因使该值减少 5 个单位。所有等位基因的频率为 0.5。（A）性状由单一基因座决定；（B）由两个基因座决定；（C）由三个基因座决定；加上少量的“随机”（环境或多基因）变异就产生了高斯曲线（D）。

型来推断基因型。

▶ 随着基因座数目的增多，性状的分布越看越像高斯曲线。加上微小的环境差异将三个基因座分布的曲线变平滑为标准的高斯曲线。

考虑到显性和变异基因的频率，再复杂的情况也会得出相同的结论。因为亲属们共有基因，所以他们的表型也是相关的，且 Fisher 在 1918 年的论文推测了不同亲属关系的亲缘程度。

在生物统计数据和多基因理论中，常误解的特征是**均值回归**（regression to the mean）。假设仅为举例，认为 IQ 变化完全是由遗传决定的。图 4.12 示，在我们简化的两个基因座模型中，对于每个 IQ 等级的母亲来说，他们子女的平均 IQ 如何为母亲 IQ 值与人群平均值的一半。这就是均值回归——但它的意义却常被误解（框 4.3）。注意这个简单的模型隐含的假设：即随机婚配。对于任一 IQ 级别的母亲，她们丈夫的平均

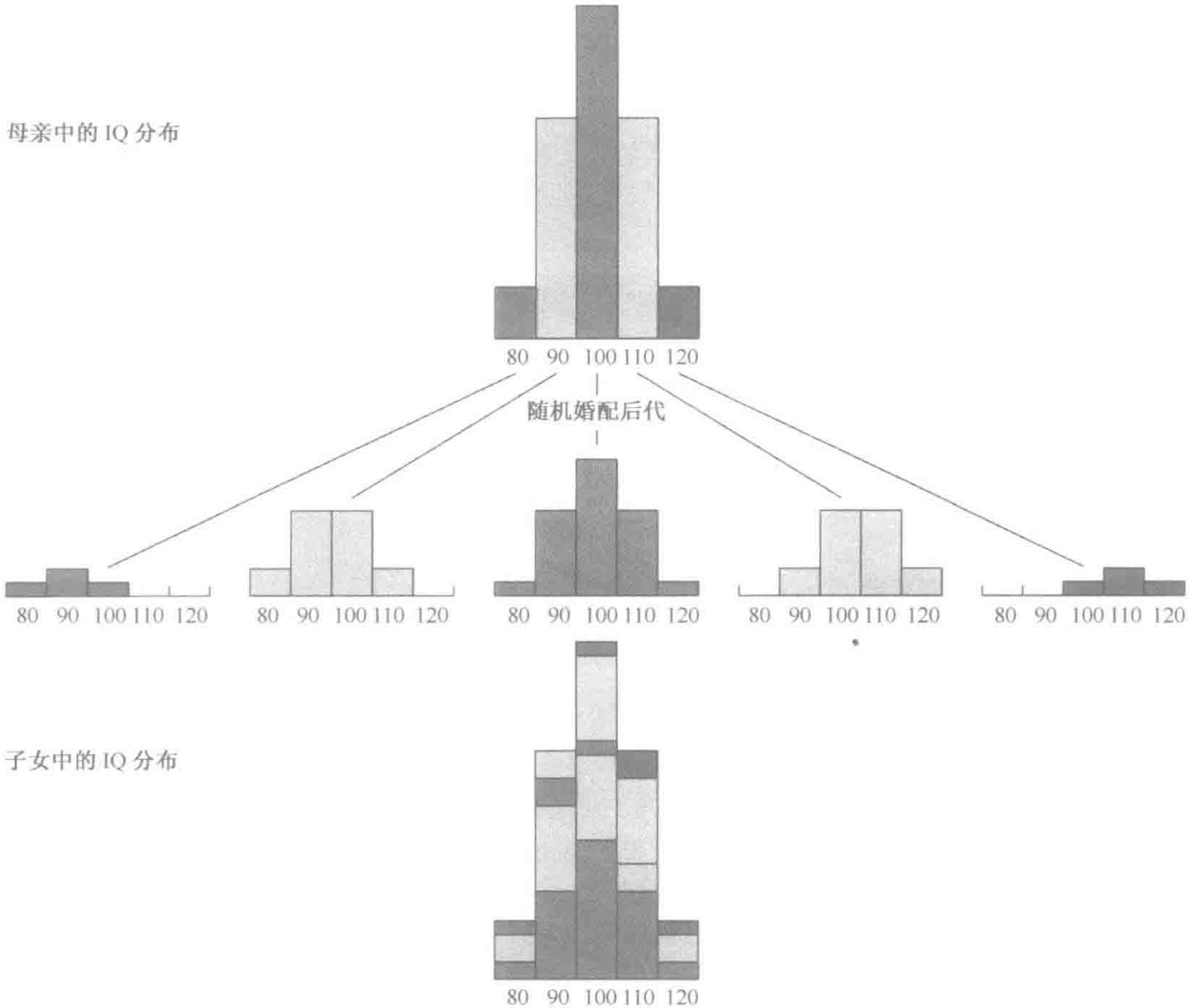


图 4.12　均值回归

与图 4.11B 相同的性状：均值为 100，由两个基因座上的共显性等位基因 A，a，B，b 决定，所有基因频率＝0.5。顶端：性状在母亲中的分布。中间：假设随机婚配，该性状在每个等级母亲的后代的分布。底部：该性状在后代中的总体分布。**注**：(a) 后代中该性状分布与在母亲的分布相同；(b) 对于每个等级的母亲，她们后代的性状均值居于母亲性状的值和人群平均值（100）之间；(c) 对于每个等级的孩子（底部），她们母亲的性状均值也是孩子性状值和人群平均值间的一半。

IQ 假设为 100。那么子女平均 IQ 就是父母 IQ 的中间值，常识也将证明这一点。现实中，高智商女性倾向于与高于平均智商的男性结婚**选择婚配**（assortative mating），即使 IQ 纯粹是遗传的性状，我们也不会预料回归到人群 IQ 均值的一半。

框 4.3　有关均值回归的两种常见的误解

（1）数代之后，每个人将恰好相同。

（2）如果某种性状表现为均值回归，它必然是遗传的。

图 4.12 所示，第一种看法是错误的。在一个简单的遗传模型中：

在每一世代中总体分布是相同的；

回归以两种方式起作用：对于每一等级的子女，他们母亲的均值是孩子的值与人群均值的一半。这听起来可能似乎矛盾，但可经调查证实，例如图 4.12 底部右手侧的柱形图（IQ120 的孩子）。他们的母亲有 1/4 IQ 为 120，1/2 为 110，1/4 为 100，使平均值为 100。

考虑第二种理解，均值回归不是遗传机制而是纯粹的统计学的现象。IQ 决定因素是否是遗传的、环境的还是二者的任意混合？如果我们选择特殊的母亲群体（如 IQ120 的），那么这些母亲一定有特殊的一套决定因素。如果我们选择共有一半决定因素的第二组（他们的子女、同胞或父母），他们的平均表型将偏离人群均值一半多。遗传学家提出的是一半的数值，但不是回归的原则。

我们简化模型的第二个假设是没有显性：每个人的表型都是每个相关基因座上每个等位基因作用的总和。如果我们考虑有显性，父母有些基因的作用将被显性等位基因的作用所掩盖而看不出其在表型中的作用，但是他们仍能传递给子女并影响其表型。如果存在显性，孩子的期望值就不再是父母的中间值了。我们对被掩盖的隐性等位基因的可能表型作用的最佳推测可通过观察其余人群获得。因此，孩子的预期表型将由父母均值趋向于群体均值来替代。替代程度取决于在表型决定中显性性状的重要程度。

遗传率是由于加性遗传效应产生方差的比例

高斯曲线由于仅有两个参数是特异的，即均值和方差（或标准差，即方差的平方根）。当方差是由独立原因引起时具有加性的有用特征（框 4.4），这样，表型的总方差 V_P 是由于各种原因引起的方差之和——环境方差 V_E 和遗传方差 V_G。V_G 又可依次分为由于单纯加性遗传效应的方差 V_A 和另外一称为显性效应的方差 V_D。性状的**遗传率**（heritability）（h^2）是整个方差中遗传方差所占的比例，即 V_G/V_P。对于对饲养高产乳量的奶牛感兴趣的农场主来说，有一饲养方案在多大程度上能使培养的畜群中奶牛的平均产奶量接近目前的最佳水平，这是一项重要的衡量指标。严格说来，V_G/V_P 为广义遗传率。显性方差不可能通过繁殖被固定，因此选择的效果是由狭义遗传率 V_A/V_P 所决定的。人类性状的遗传率通常被评价为分离分析的一部分（节 15.2 和表 15.4）。然而切记，对许多人类行为特征，将方差简单划分为环境的和遗传部分是不适用的。我们既给了孩子基因，也给了他们环境。遗传的和社会的不利条件总是共同伴随的，故遗传因素和环境因素通常是相互关联的。如果遗传和环境因素不是独立的，V_P 就不等于 V_G+V_E，有加性效应方差。方差的增大能迅速削弱此模型说明的力度，总的来说这已是一个难以研究的领域。

“遗传率”一词常常被人误解。遗传率与遗传方式完全不同。遗传方式（常染色体显性，多基因遗传等）是某性状的固有特征，但遗传率却不是。“IQ 的遗传率”是“IQ 方差的遗传率”的缩略，比较这两个问题：

- IQ 多大程度上是遗传性的？这是个没有意义的问题。
- 在特定地区和特定时间，个体间 IQ 差异有多大程度是由遗传差异引起的，又有多大程度是由不同的环境和生活史引起的？这即使难以回答，但是个有意义的问题。

在不同的社会环境中，IQ 的遗传率将有所不同。社会越平等，IQ 的遗传率就应该越高。如果每个人拥有相同的机会，那么人们之间的一些环境差异就被消除了。因此继续存在的 IQ 差异则更多是由人们之间的遗传差异造成的。

框 4.4　方差的划分

表型方差（V_P）＝遗传方差（V_G）＋环境方差（V_E）

V_G＝加性遗传效应方差（V_A）＋显性效应方差（V_D）

$V_P = V_A + V_D + V_E$

遗传率（广义）＝ V_G / V_P

遗传率（狭义）＝ V_A / V_P

4.4.3　不连续性状的多基因理论

像身高、体重等大多数经典的“多基因”连续变异性状不会引起医学遗传学家多大兴趣（尽管我们讨论了肥胖：节 15.6.8）。他们更多的兴趣是在于数不清的倾向于家族中传递的疾病和畸形上，但这些疾病不表现为孟德尔系谱方式。非孟德尔遗传的主要概念工具是 Falconer 提出的把多基因理论扩展为双歧或非连续性状（那些或有或无的性状）。

Falconer 推测潜在的连续变量的**易患性**（susceptibility）。你可患也可不患腭裂，但每个胚胎都有一定患腭裂的易患性。易患性可低可高：它是多基因性的，在人群中呈高斯分布。Falconer 还推测**阈值**（threshold）的存在。易患性超过临界阈值的胚胎发育为腭裂；而那些易感性低于阈值的胚胎，即使是刚刚低于阈值也能避免腭裂发生。抛开数学的精妙，此模型正如图 4.13 所示。可以把阈值想像成天平的中立点。改变因素的平衡表型就会倾向于这端或那端。

对于腭裂，多基因阈值模型直观上似乎是合理的（Fraser，1980）。所有的胚胎开始有腭裂。在早期发育过程中，腭突板必须变平、彼此融合。这些事件必须发生在特定的发育窗口时期。众多不同遗传的和环境因子影响胚胎的发育，因此，易患性应该是多基因的似乎很有道理。腭突板是否相交、有足够的时间融合，或它们恰恰及时融合并不重要——如果融合了就是正常的腭突类型；若是没有融合，就成了腭裂。因此，有一个自然的阈值附加在连续可变的过程。

Falconer 的阈值理论有助于解释家族中再发风险如何变化。受累个体必然有高易感性等位基因的不幸组合。一般来说与他们共有基因的亲属的易感性增高，其与人群均值的分散度依赖于携带共有基因的比例。因此，多基因阈值性状趋向于在家族中传递（图 4.14）。有几个受累患儿的父母可能恰恰是不幸的，但平均来讲他们比只有一个患儿的父母具有更多的高风险等位基因。阈值是固定的，但平均易患性和由此而来的再发风险

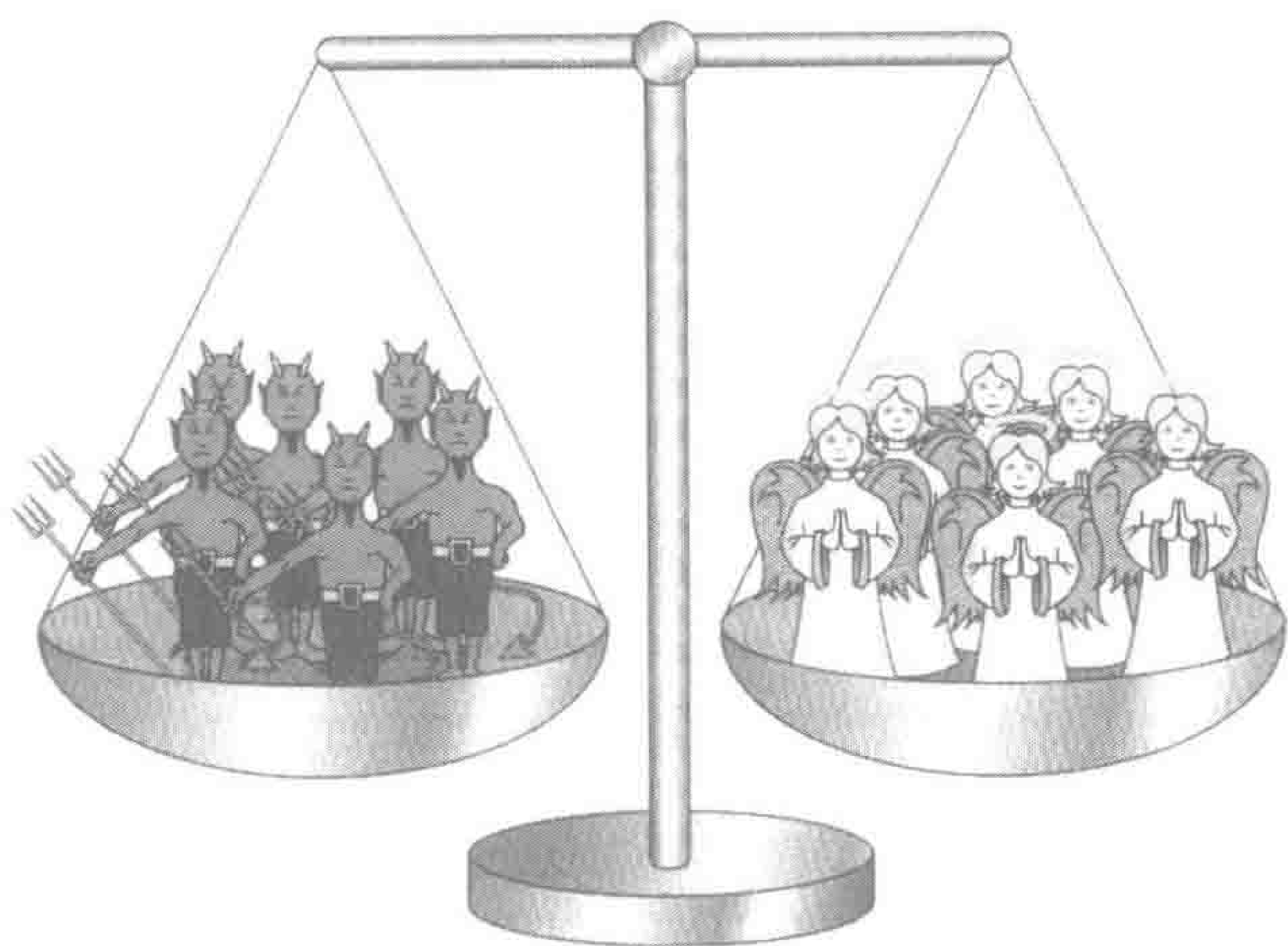


图 4.13　由多因子决定的疾病或畸形

天使和魔鬼代表遗传和环境因素的任意结合。不考虑致病的特殊因素，额外增加一个魔鬼或减少一个天使都会使平衡倾斜。R. S. W Smithells 教授赠图。

随着以前受累患儿数目增多而提高。

许多假定了阈值的疾病在两种性别中具有不同的发病率，意指性别特异性阈值。例如先天性幽门狭窄，通常男孩比女孩多 5 倍。女孩的阈值一定高于男孩，因此受累女孩的亲属比受累男孩的亲属具有更高的易患性（图 4.15）。虽然在每个病例中假如是男孩，该婴儿受累风险比女孩高 5 倍，但是受累女孩亲属的再发风险相应更高（表 4.1）。

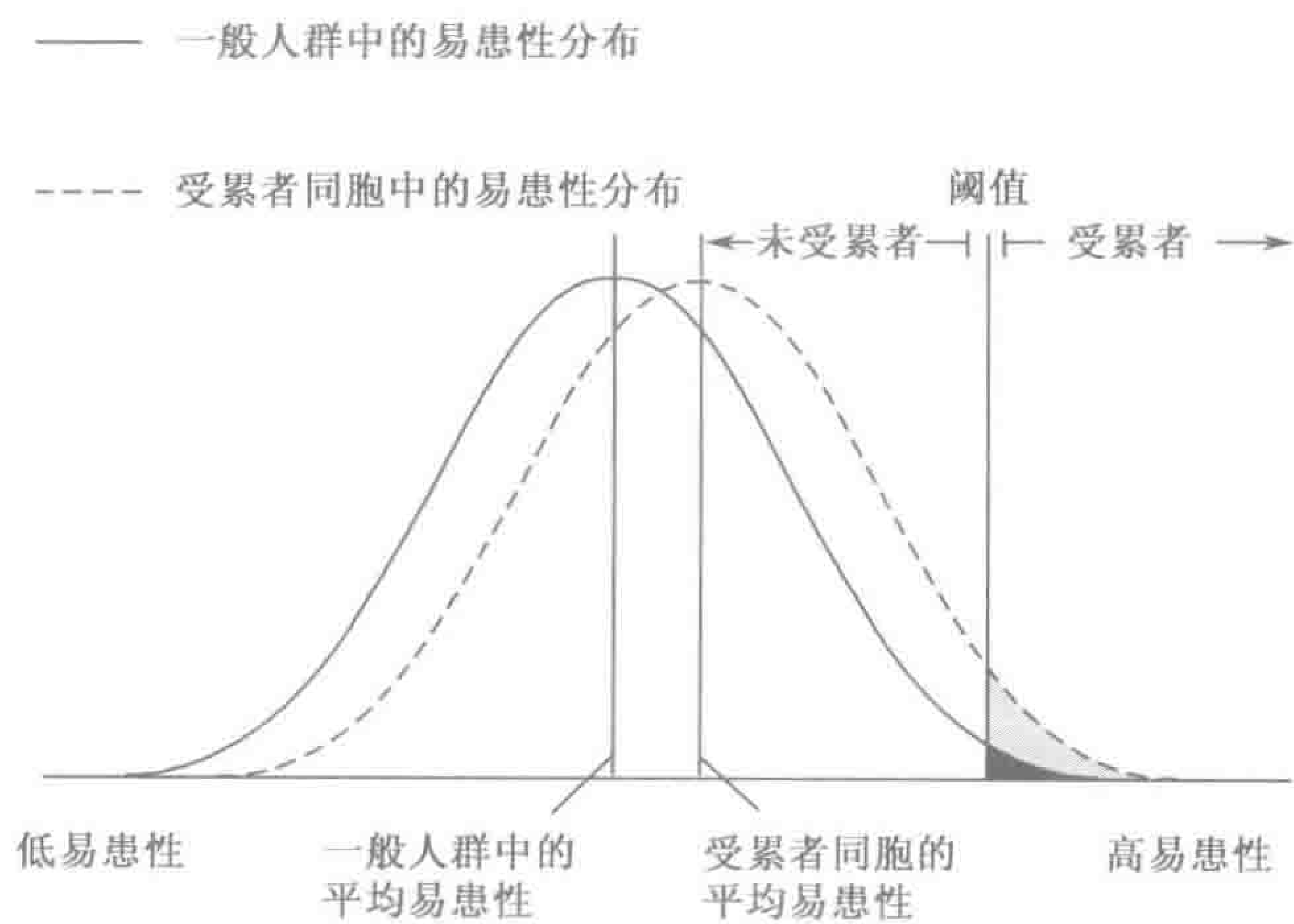


图 4.14　Falconer 的非孟德尔双歧性状的多基因阈值模型

性状易患性是多基因的，呈常态分布（实线曲线）。易患性高于一定阈值的个体（图 4.13 中的平衡点）患病。他们的同胞（虚线曲线）比人群平均易患性高，其中大部分易患性超出了阈值。因此该疾病倾向于在家族中传递。

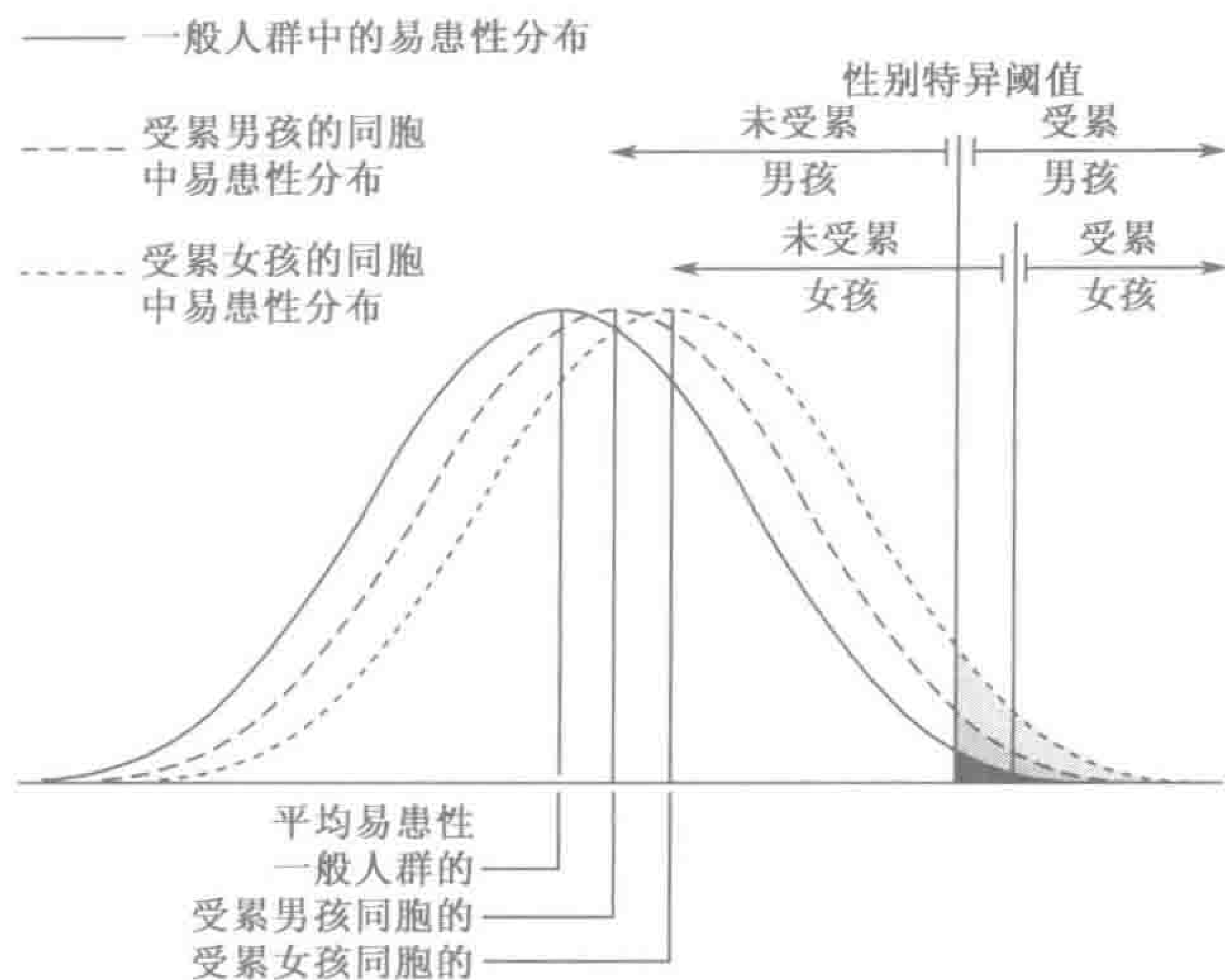


图 4.15　具有性别特异阈值的多基因性状

如果一个非孟德尔双歧性状主要影响男性，通过推测男性阈值比女性要低，这与多基因阈值理论相符合。由此得出受累女性的亲属再发风险更高，但是那些再发病例会是男性。见表 4.1 中符合此解释的数据例子。

表 4.1　幽门狭窄的再发风险

亲 属	儿子	女儿	兄弟	姐妹
男性先证者	19/296	7/274	5/230	5/242
	(6.42%)	(2.55%)	(2.17%)	(2.07%)
女性先证者	14/61	7/62	11/101	9/101
	(22.95%)	(11.48%)	(10.89%)	(8.91%)

受累男孩多于女孩，但是受累女孩亲属再发风险更高。此数据符合性别特异性阈值理论模型的多基因阈值（图 4.15）。数据引自 Fuhrmann 和 Vogel（1976）。

4.4.4　非孟德尔疾病的咨询运用经验风险率

在非孟德尔疾病遗传咨询时，风险值不是从多基因理论得来；而是如表 4.1 通过对人群的调查得到的**经验风险率**（empiric risk）。这与孟德尔性状的 1/2、1/4 等源于理论的风险值完全不同。家族史的影响也不尽相同。如果一对夫妇双方都是囊性纤维化的携带者，他们下一代患病风险为 1/4。不管他们已经生育了几个患病的或正常的孩子，风险率不变（概括在标签“风险无记忆”中）。如果他们生出一神经管缺陷的婴儿，其再发风险在英国约为 1/25，但如果他们已经有了两个受累患儿，再发风险率为 1/12（也许由“已给予他的风险”来总结）。并非生了第二个男患儿能使再发风险增加，而是让我们认识到他们是一直处于特殊高风险的一对夫妇。有好讽刺的人会说咨询师是事后诸葛亮。但是在不完善的知识背景下，实践与我们基于对阈值理论的理解以及与流行病学数据的一致，代表了我们能提供的最好咨询。

4.5 影响基因频率的因素

4.5.1 基因频率和基因型频率之间可有的简单关系

一个假想的实验：从基因库中提取基因

整个人群中，尽管每个人在某一特定基因座上只有两个相同或不同的等位基因，但这个基因座可能会有多种不同的等位基因。我们可假设在人群中 A 基因座上所有的等位基因组成的一个**基因库**（gene pool）。等位基因 A_1 的**基因频率**（gene frequency）是基因库中 A_1 占所有 A 等位基因的比例。假设在 A 基因座上有两个等位基因 A_1、A_2。基因频率分别为 p、q（p、q 介于 0 到 1 之间）。让我们做个假定的实验：

- ▶ 从基因库中随机取出一个等位基因，A_1 的几率是 p，A_2 的几率是 q。
- ▶ 随机取出第二个等位基因，再次得到 A_1 的几率是 p，为 A_2 的几率是 q（假设基因库足够大，取出第一个等位基因对基因库中的基因频率没有大的改变）。结果显示：
  - 两等位基因均为 A_1 的几率是 p^2；
  - 两等位基因均为 A_2 的几率是 q^2；
  - 第一个等位基因为 A_1、第二个等位基因为 A_2 的几率是 pq。第一个等位基因为 A_2、第二个等位基因为 A_1 的几率是 qp。总体看来，得到一个 A_1 等位基因、一个 A_2 等位基因的几率为 2pq。

Hardy-Weinberg 分布

如果我们从人群中随机挑选出一个人，这与随机从基因库中挑选出两个基因是等价的。此人为 A_1A_1 的几率为 p^2，为 A_1A_2 的几率为 2pq，为 A_2A_2 的几率是 q^2。无论何时从基因库独立随机地选取两个基因，这种基因频率和基因型频率间的简单关系（Hardy-Weinberg 分布见框 4.5）都成立。A_1 和 A_2 可以是基因座上仅有的等位基因（例子中 $p+q=1$）或是有其他等位基因和其他基因型（$p+q<1$）。对于 X 连锁基因座的男性，作为半合子（只有一个等位基因）是 A_1 或 A_2，其频率分别是 p、q，而女性可能是 A_1A_1、A_1A_2、A_2A_2（框 4.5）。

框 4.5 Hardy-Weinberg 平衡：等位基因频率 p（A_1）和 q（A_2）的基因型频率

	常染色体基因座			X 连锁基因座				
				男 性		女 性		
基因型	A_1A_1	A_1A_2	A_2A_2	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
频　率	p^2	2pq	q^2	p	q	p^2	2pq	q^2

注：无论在此基因座上是否是 A_1、A_2 仅有的等位基因，都将得到这些基因型频率。

Hardy-Weinberg 分布的局限性

如果违背了某个体的两个基因是从基因库中独立抽取的潜在假设，那这些简单计算

关系就不成立了。是否为随机婚配是问题的关键所在。选择婚配有几种形式，但通常最为重要的是**近亲婚配**（inbreeding）。如果你和你的亲属结婚，那么你是在和与你自己基因相似的人结婚。这就增加你们孩子是纯合子的可能性，而降低他们是杂合子的可能性。罕见的隐性性状与父母近亲婚配是强烈相关的，忽略了近亲婚配的 Hardy-Weinberg 算法将高估整个人群中携带者的频率。

在遗传咨询中 Hardy-Weinberg 分布的用途

对多种形式的遗传分析，诸如连锁分析（节 13.3）、分离分析（节 15.2），基因频率或基因型频率都是必不可少的，他们在计算遗传风险时有特殊的价值。框 4.6 举例说明。

框 4.6　Hardy-Weinberg 分布可用于（慎用）计算携带者频率及咨询简单的发病风险

一个常染色体隐性疾病 10 000 人中有 1 个受累。那么携带者的预期频率是多少？

表　型：	正常		受累
基因型：	AA	Aa	aa
频　率：	p^2	$2pq$	$q^2=1/10000$

q^2 为 10^{-4}，因此 $q=10^{-2}$ 或 1/100。

A 基因座上 100 个基因中有 1 个是 a，99/100 是 A。

携带者频率 2pq，就是 2×99/100×1/100，非常近似 1/50。这说明该性状频率近亲婚配后未增高。

如果上述疾病患儿的双亲再婚，新婚姻中再生出患儿的风险是多少？

若生出患儿，父母双方必定是携带者，于是风险为 1/4，因此总体风险是：

（双亲是携带者的风险）×（新配偶为携带者的风险）×1/4

=1×1/50×1/4

=1/200

这是假设新配偶的家族中无相同疾病的家族史的情况下。

X 连锁的红绿色盲，在英国 1/12 的男性受累。那么女性是携带者的比例是多少？

受累女性的比例是多少？

	男性		女性		
基因型：	A_1	A_2	A_1A_1	A_1A_2	A_2A_2
频　率：	p	$q=1/12$	p^2	$2pq$	q^2

q=1/12，因此 p=11/12

2pq=2×1/12×11/12=22/144

$q^2=1/144$。故这个单一基因座模型预测，15%女性将是携带者，0.7%女性受累。

4.5.2　基因型频率可用于（但要慎用）计算突变率

突变基因由新突变产生、被自然选择淘汰（框 11.2）。对于既定的选择水平，我们能计算出用来替代由于选择而丢失的基因的突变率。假定人群中的基因丢失频率和替代频率处于平衡，通过计算我们就可知目前的突变率。我们将**选择系数**（coefficient of

selection）（s）定义为由于选择，一个基因型不能繁衍的相对几率（群体最适合生存类型的 s=0，遗传致死性 s=1）。

- 对于**常染色体隐性**（autosomal recessive）疾病，人群中 q^2 比例的人受累。每代致病基因丢失的比例为 sq^2。突变率为 μ（$1-q^2$）则平衡，这里 μ 为每个基因每代的突变率。平衡时 $sq^2=\mu(1-q^2)$，或近似 $\mu=sq^2$（如果 q 很小）。
- 对于**常染色体显性**（autosomal dominant）疾病，纯合子相当少见。杂合子发生频率为 2pq（致病基因频率=p）。只有一半基因由于不能生育而丢失的为致病等位基因。因此基因丢失比例非常接近 sp。这将被新突变率 μq^2 所平衡，如果 q 接近为 1，μq^2 近似为 μ。故 $\mu=sp$。
- 对于 **X-连锁隐性**（X-linked recessive）疾病，通过受累男性丢失基因的频率为 sq。由于人群中所有的 X 染色体都可能突变，但仅男性 1/3 的 X 染色体要面临选择，丢失基因的频率被 3μ 的突变率所平衡，故 $\mu=sq/3$。

这些结果总结在框 4.7 中。从许多有机体的研究应用以上结论做出的估计可与一般性的推测比较，每个基因每代的突变率一般为 $10^{-5}\sim10^{-7}$。

框 4.7　突变-选择平衡

常染色体隐性疾病*	$\mu=sq^2$	or	$\mu=F(1-f)$
常染色体显性疾病	$\mu=sp$	or	$\mu=1/2F(1-f)$
X-连锁隐性疾病	$\mu=sq/3$	or	$\mu=1/3F(1-f)$

μ=每代每个基因的突变率

p，q=基因频率

s=选择系数

f=生物适合度=1−s

F=人群中该性状的频率

* 如果存在杂合子优势，则此公式的突变率预测值将产生严重错误，见框 4.8。

4.5.3　在确定隐性疾病频率时杂合子优势比再发突变更重要

公式 $\mu=sq^2$ 给出常染色体隐性疾病意想不到的高突变率。以囊性纤维化（CF）为例。直到最近，患有 CF 的人无人能存活到生育年龄，因此 s=1。在英国大约为出生人口的 1/2000 患 CF，即 $q^2=1/2000$，由公式得出 $\mu=5\times10^{-4}$，这对任何基因来说，是相当高的突变率——但是有证据表明 CF 新突变实际上很罕见。这是从 CF 的不平衡种族分布和严重连锁不平衡的存在得出的（节 13.5.2）。

忽略的因素是**杂合子优势**（heterozygote advantage）。CF 携带者具有或曾有超过正常纯合子的生育优势。对这种优势可能是什么一直有争议。CF 基因编码细胞膜氯通道，为伤寒沙门杆菌进入上皮细胞所需要。因此可能杂合子抵抗伤寒热的能力相对强（Pier *et al.*，1998）。无论杂合子优势的原因何在，如果 s_1 和 s_2 分别是 AA、aa 基因型的选择系数，当 A 和 a 的基因频率的比例 p/q 为 s_2/s_1 时就建立了平衡。框 4.8 列举了 CF 的计算结果，显示杂合子优势小到在人群调查中不易观察到，但可能对基因频率变

化起主要影响。

框 4.8　有利于 CF 杂合子的选择

对于 CF，在英国疾病频率约为出生人口的 1/2000。

表　型：	未受累		受累
基因型：	AA	Aa	aa
频　率：	p^2	$2pq$	$q^2=1/2000$

q^2 为 5×10^{-4}，因此 $q=0.022$ $p=1-q=0.978$

$p/q=0.978/0.022=44.45=s_2/s_1$

如果 $s_2=1$（受累纯合子不能生育），$s_1=0.022$

如果 Aa 杂合子比 AA 纯合子后代平均存活率高 2.2%，即使没有新的突变，现有的 CF 基因频率将会维持下去。

值得记住的是，医学上重要的孟德尔疾病是那些常见而又严重的疾病。它们必定有这样那样的特殊窍门才能在面临选择保持常见。这个窍门可能是格外高的突变率（杜兴肌营养不良），或非病理性前突变的传代（脆性 X 综合征），或过了生育年龄延迟发病（亨廷顿病）——但是，对于常见的严重隐性遗传病来说，最通常的诀窍是杂合子优势。

（陈芳杰　译）

进一步阅读

Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Longman, Harlow.

Forrest DW (1974) *Francis Galton: The Life and Work of a Victorian Genius*. Elek, London.

参考文献

Fisher RA (1918). The correlation between relatives under the supposition of mendelian inheritance. *Trans. Roy. Soc.* **52**, 399–433.

Fuhrmann W, Vogel F (1976) *Genetic Counselling*. Springer, New York.

Fraser FC (1980) The William Allan Memorial Award Address: Evolution of a palatable multifactorial threshold model. *Am. J. Hum. Genet.* **32**, 796–813.

Harper PS (2001) *Practical Genetic Counselling*, 5th Edn. Arnold, London.

Heutink P, van der Mey AG, Sandkuijl LA *et al.* (1992) A gene subject to genomic imprinting and responsible for hereditary paragangliomas maps to chromosome 11q23-qter. *Hum. Molec. Genet.* **1**, 7–10.

Jobling MA, Tyler-Smith C (2000) New uses for new haplotypes: the human Y chromosome, disease and selection. *Trends Genet.* **16**, 356–362.

Pier GB, Grout M, Zaidi T *et al.* (1998) *Salmonella typhi* uses CFTR to enter intestinal epithelial cells. *Nature* **393**, 79–82.

Prezant TR, Agapian JV, Bowman MC *et al.* (1993) Mitochondrial ribosomal RNA mutation associated with both antibiotic induced and non syndromic deafness. *Nature Genet.* **4**, 289–294.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837.

Strain L, Dean JCS, Hamilton MPR, Bonthron DT (1998) A true hermaphrodite chimera resulting from embryo amalgamation after in vitro fertilization. *N. Engl J. Med.* **338**, 166–169.

Van der Meulen MA, van der Meulen MJP, te Meerman GJ (1995) Recurrence risk for germinal mosaics revisited. *J. Med. Genet.* **32**, 102–104.

Viljoen D, Ramesar R (1992) Evidence for paternal imprinting in familial Beckwith–Wiedemann syndrome. *J. Med. Genet.* **29**, 221–225.

Wilkie AOM (1994). The molecular basis of dominance. *J. Med. Genet.* **31**, 89–98.

第5章　扩增DNA：PCR和细胞DNA克隆

本章内容

5.1　DNA克隆的重要性

目前DNA技术的基础在很大程度上是基于两种完全不同的方法来研究复杂DNA群体中的特异DNA序列（图5.1）：

- **DNA克隆**（DNA cloning）　必须有选择性地扩增目的DNA序列或片段以制备大量相同的拷贝，使目的产物得以纯化。之后，DNA的结构和功能得以广泛深入地研究。
- **分子杂交**（molecular hybridization）　目的片段没有以任何方式被扩增或纯化；取而代之，在含有许多不同序列的复杂混合物中被特异检测。

在DNA克隆以前，我们对DNA的了解非常有限。DNA克隆技术改变了这种状况，使遗传学研究发生了变革。究其原因，有必要了解DNA序列的惊人大小和复杂性（与蛋白质序列比较而言）。个体的核DNA分子包含上亿个核苷酸。当利用标准方法从细胞中分离DNA时，这些巨大分子在剪切力作用下变成片段，产生了DNA片段仍然很大的复杂混合物（典型的长度是50～100kb）。

考虑到从典型的真核细胞，或甚至原核细胞中提取的DNA的复杂性，面临的挑战就是如何分析它。我们所需要的是可以将DNA分开的某种方法，这样DNA序列的不同亚群能够得以纯化。对许多真核细胞而言，一种早期研究方法是通过离心技术分离

DNA 序列的不同群体。平衡密度梯度的超速离心技术（例如，氯化铯密度梯度）通常将真核细胞 DNA 分离成一个主带（大体积 DNA）和几个次带。次带 DNA 的浮力密度与大体积 DNA 不同（且彼此亦不同），因为它们是由随机重复的**卫星 DNA**（satellite DNA）序列组成，后者的碱基组成明显不同于大体积 DNA。已经发现卫星 DNA 与染色体的特异结构和特异功能有关。尽管这是具有价值和有意义的，但纯化的卫星 DNA 只是基因组的较少成分，且不含有基因。DNA 克隆提供的是一种纯化和研究任何 DNA 序列的普通方法。

DNA 克隆要求选择性扩增大而复杂 DNA 群体中的特异 DNA 成分（靶 DNA），可以是一种特殊组织（或细胞类型）中的总基因组 DNA，或由特殊组织总 RNA 制备的**互补 DNA**（complementary DNA，cDNA）。利用 DNA 聚合酶，可以在体外或细胞内进行扩增。

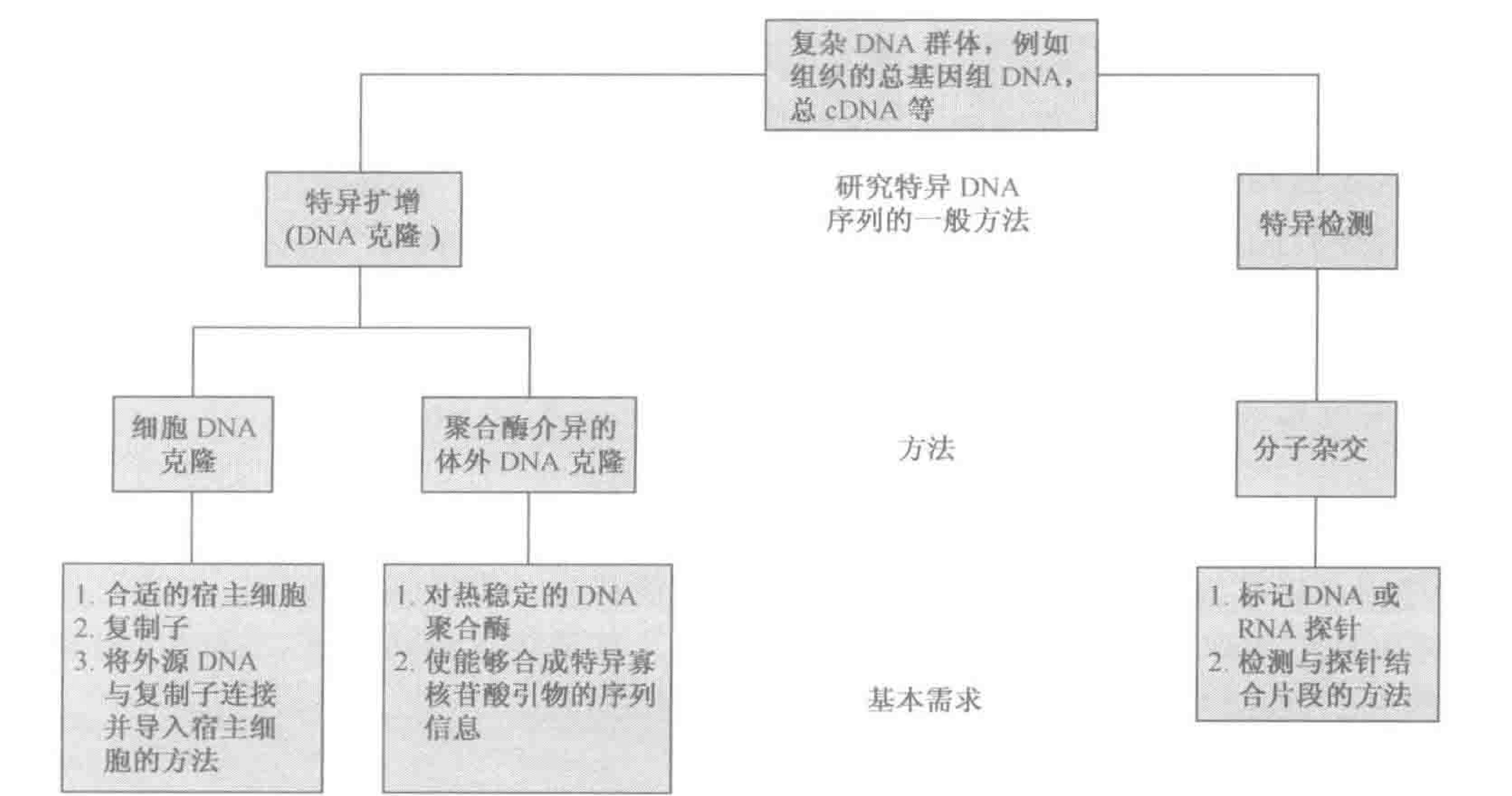


图 5.1　在复杂 DNA 群体中研究特异 DNA 序列的一般方法

体外 DNA 克隆

这里，利用能够与靶 DNA 特异结合的寡核苷酸引物筛选目的靶 DNA。一旦序列-特异引物与靶 DNA 结合，热稳定的 DNA 聚合酶可以产生靶序列的附加拷贝。靶序列的新 DNA 拷贝仍然在称为**聚合酶链反应**(PCR）的链反应 polymerase chain reaction，中交替充当模板，制备更多的拷贝。

细胞 DNA 克隆

靶 DNA 分子被连接到单一类型**复制子**（replicon）分子上（在合适的宿主细胞内能够进行独立的 DNA 复制），并且被导入合适的宿主细胞。靶-复制子的杂交分子典型地经历多次 DNA 复制循环（**DNA 扩增**，DNA amplification）之后，被扩增的靶 DNA

分子可以从细胞的其他成分，也可以从复制子分子中分离。细胞 DNA 克隆的关键是当不同靶-复制子分子被导入细胞（**转化**，transformation）时，典型地，每个细胞仅接纳一种类型的靶 DNA -复制子分子，所以细胞的所有子代准确地含有同种类型的靶 DNA 分子（**DNA 克隆**，DNA clone）。通过从被分离的细胞集落开始生长大量细胞，并可能制备纯的目的靶 DNA 群体。

5.2 PCR：基本特征和应用

借助于 DNA 的迅速克隆和分析，PCR 已经使分子遗传学发生了变革。自从 1980 年首次报道描述了这一新技术以来，PCR 技术已经广泛地应用于基础和临床研究。

5.2.1 PCR 和反转录酶（RT）PCR 的基本原理

起始核酸的选择

通常，PCR 被设计在异源性 DNA 序列的集合体内选择性扩增特异靶 DNA 序列。起始 DNA 经常是某一特殊组织或培养细胞的总基因组 DNA，在那种情况下，靶 DNA 通常是起始 DNA 的极小部分。例如，如果打算扩增人类 DNA（二倍体基因组大小＝3300Mb）中 1.6kb 的β珠蛋白基因，靶 DNA：起始 DNA 比率是 1.6kb：3300Mb 或 1：2 000 000。许多 PCR 反应可以扩增甚至更小的靶 DNA，诸如平均 140bp 的单个外显子，所以相当于不到人类基因组 DNA 一起始群体的 0.000005％。

就编码 DNA 靶序列而言，起始 DNA 经常可能是总 cDNA，cDNA 是通过合适的组织或细胞系提取 RNA，然后利用反转录酶（RT-PCR）将 RNA 转化成 cDNA。取决于靶序列在原来 RNA 群体中作为转录物被表达的程度，靶序列的数量可能显著增加（当与基因组 DNA 该序列的代表比较时）。然而，在最优化的条件下，有时可以利用 RT-PCR 扩增仅仅含有基础转录水平的组织中的靶序列。

引物的选择

为了进行选择性扩增，需要靶序列的某些更重要的 DNA 序列信息。这些信息被用于设计两个寡核苷酸引物（**扩增引物**，amplimer），最好是 18～25 个核苷酸长，对靶序列的旁侧序列是特异的（即引物的碱基序列完全代表靶序列的旁侧序列）。

就大多数 PCR 反应而言，目的是扩增单一 DNA 序列，所以除了靶位点外，减少引物与其他位点的结合机会是重要的。因此，重要的是避免 DNA 重复序列，并且引物设计需要考虑各种其他因素：

- 碱基组成。GC 含量应该在 40％～60％之间，四种核苷酸尽可能随机分布；
- 融解温度（Tm，见节 6.2.1 定义）。计算的两个引物的总 Tm 差异不应该＞5℃，扩增产物的 Tm 与引物的 Tm 相比，差异不应该＞10℃；
- 3′端序列。一个引物的 3′ 端序列不应该与同一反应中其他引物的任何部分序列互补。注：引物 3′端碱基的正确配对是至关重要的，可以用来确保不是其他的而是甚至一个特异等位基因的扩增，以此奠定了**等位基因特异 PCR**（allelespecific PCR）的基础（框 5.1）；

▶ 自身互补序列 避免反向重复序列或任何自身互补序列长度＞3bp。

各种商业化软件程序和免费商品程序如在http：//www. hgmp. mrc. ac. uk/GenomeWeb/nucprimer. html 上编辑的那些程序有助于引物设计。

框 5.1　PCR 方法的词汇表

▶ **等位基因特异性 PCR**（allele-specific PCR）　设计扩增一 DNA 序列同时排除扩增其他等位基因的可能性。要求 PCR 引物 3′端和靶 DNA 之间碱基精确配对见图 5.4。

▶ Alu-PCR　利用与 Alu 重复序列特异的引物进行的 PCR 方法，人类基因组 DNA，大约每 3kb 出现一次 Alu 重复序列。当相邻 Alu 重复序列方向相反时，单一类型 Alu 引物可以扩增间隔序列（例如图 A 中的引物 A 和引物 B 方向相反）。

▶ **锚定 PCR**（ anchored PCR ）　利用序列特异引物和通用引物扩增已知序列的邻近序列。通用引物识别并与通用序列结合，后者是人工添加到起始群体的所有不同 DNA 分子上，例如通过共价连接到双链**寡核苷酸接头**（oligonucleotide linker）上。

▶ **差异展示 PCR**（ differential display-PCR ）　比较两个相关细胞来源表达的 mRNA 群体的一种 RT-PCR 形式，目的是找到差异表达的基因（图 7.16）。

▶ DOP-PCR（变性寡核苷酸引导的 PCR）使用部分变性寡核苷酸引物（degenerate oligonucleotide primer）（合成几套寡核苷酸序列，这些序列在某些核苷酸位置有相同的碱基，而在其他位置不同）扩增各种相关靶 DNA。

▶ **热启动 PCR**（ hot-start PCR ）　增加 PCR 反应特异性的一种方法。开始热变性步骤之前混合所有 PCR 试剂，可以使引物序列非特异结合的机会增多。为了减少这种可能性，PCR 的一个或多个成分自然分开，直到第一次变性步骤完成。

▶ **反向 PCR**（ inverse PCR ）　一种获取与已知 DNA 序列最邻近 DNA 的方法（参看锚定 PCR）。在这种情况下，用限制性核酸酶消化起始 DNA 群体，稀释成低 DNA 浓度，然后，用 DNA 连接酶处理，通过分子内连接促进环状 DNA 分子形成。PCR 引物被定位以便与已知 DNA 序列结合，然后启动新 DNA 合成，合成方向背离已知序列，朝向未知邻近序列，引起未知序列的扩增。见图 B 的对面（X 和 Y 是已知序列旁侧未鉴定序列）。

▶ **岛屿-营救 PCR**（ island-rescue PCR ）　在 CpG 岛扩增序列的特异方法。

▶ **接头引导 PCR**（ linker-primed PCR ）**连接适配子 PCR**（ ligation adaptor PCR）　不加选择的扩增形式。用限制性核酸酶消化复杂的起始 DNA，产生带有相同类型突出末端的多个片段。带有相似突出末端的已知序列的双链寡核苷酸接头被连接到片段上。然后，利用与接头序列特异的引物扩增接头分子旁侧的所有 DNA 片段进行 PCR 反应。

▶ **巢式引物 PCR**（ nested primer PCR ）　一种增加 PCR 反应特异性的方法。稀释起始扩增反应产物，作为第二轮反应的起始 DNA 模板。第二轮反应使用一套不同的引物。这对引物对应的序列与第一次反应的引物接近，但是在其内。

▶ **定量 PCR**（ quantitative PCR ）　见**实时 PCR**（real-time PCR）。

▶ RACE-PCR　快速扩增 cDNA 末端（见图 7.12）的一种**锚定-引导 PCR**（anchor-primed PCR）（见上文）。

▶ **实时 PCR**（ real-time PCR ）　利用荧光检测热循环仪扩增特异核苷酸序列并且同时检测其浓度的一种定量 PCR 方法。主要有两种研究应用：（Ⅰ）量化基因表达（通过微阵杂交分析证实被检测的基因差异表达）；并且（Ⅱ）筛查突变和单核苷酸多态性。在分析实验室也可用于检测临床和工业样本中 DNA 或 RNA 序列的丰度。

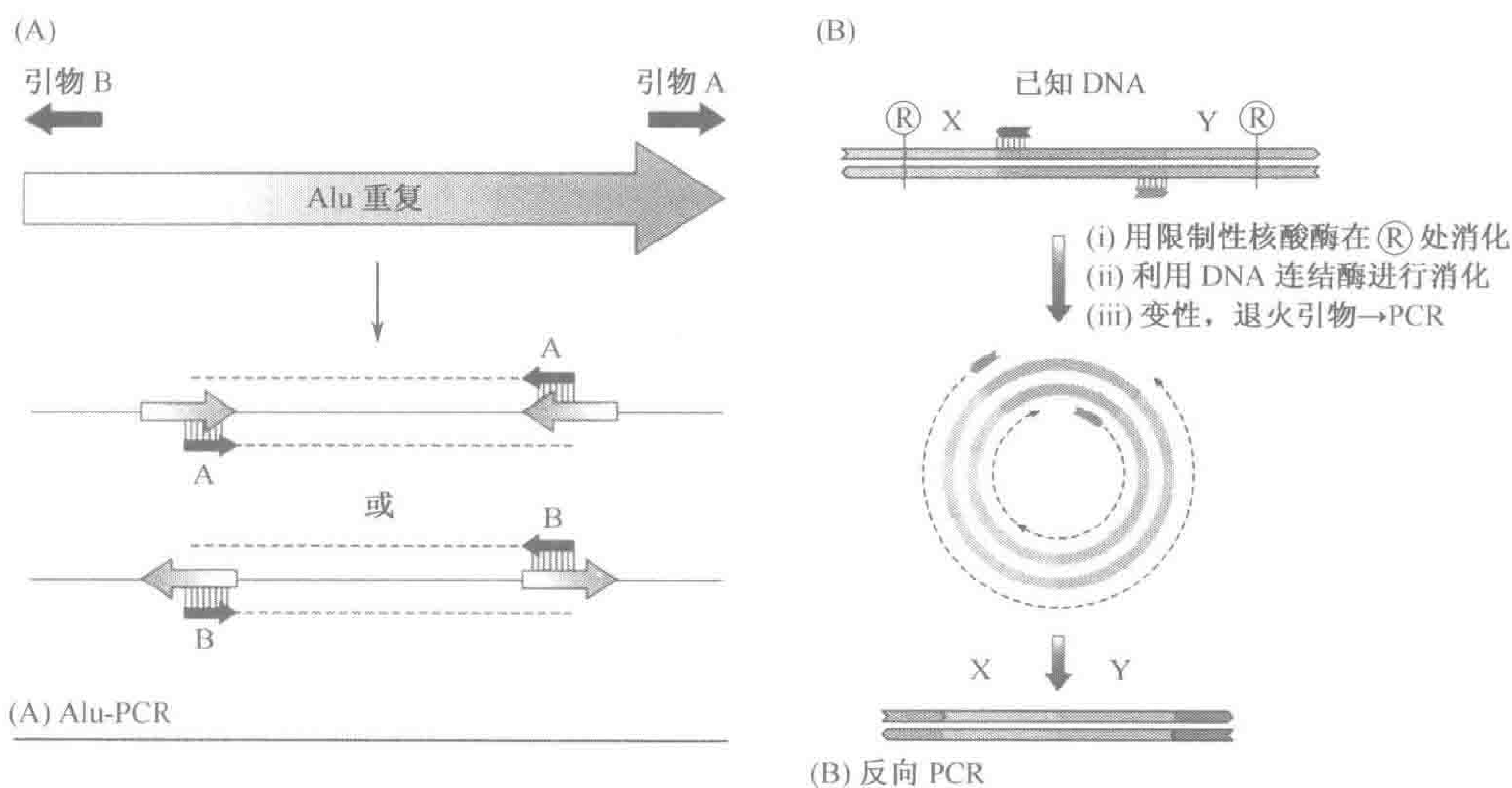
▶ **RT-PCR（反转录酶 PCR）**（ reverse transcriptase PCR ）　要求起始群体是 mRNA 的 PCR 且起始反转录酶步骤产生 cDNA。

框 5.1 PCR 方法的词汇表（续）

▶ **降落 PCR**（touch-down PCR） 增加 PCR 反应特异性的一种方法。大多数热循环仪可以程序化完成 PCR 循环。其中复性温度从超过预期 Tm 的始动值逐渐降低到 Tm 以下。通过保证最初杂交的高度严格性，减少非特异产物形成，使预期序列占优势。

▶ **全基因组 PCR**（whole genome PCR） 不加选择 PCR。使用广泛变性的引物或将寡核苷酸接头连接到一复杂 DNA 群体，然后用接头特异寡核苷酸引物扩增所有序列。

注：通过设计与序列互补的两个核苷酸构建双链寡核苷酸接头（**适配子**，adaptor），在独立的化学反应中分别合成两个核苷酸，且一旦纯化，它们可以相互碱基配对形成目的双链序列。寡核苷酸接头与 DNA 序列连接可以进行：（Ⅰ）利用与接头特异的寡核苷酸的 PCR；或（Ⅱ）加入想要的特性，例如，有助于克隆的限制性位点（图 5.10）。含有几个限制性位点的复杂多接头（polylinker）常规插入克隆载体（节 5.3.5）。



(A) Alu-PCR

(B) 反向 PCR

循环特征和指数扩增

PCR 由三个连续反应的一系列循环组成：

▶ **变性**（denaturation）人类基因组 DNA 变性典型地开始在大约 93～95℃；

▶ **引物复性**（primer annealing）温度通常从 50～70℃，取决于预期双链的熔解温度（再次复性温度典型地开始在低于计算的熔解温度 5℃左右）；

▶ **DNA 合成**（DNA synthesis）典型地约在 70～75℃。使用的 DNA 聚合酶对热稳定（此酶需要在 70～75℃有效地延伸且不应受到变性步骤的副作用影响）。在合适的对热稳定的 DNA 聚合酶和 DNA 原料（四种三磷酸脱氧核糖核苷，dATP，dGTP，dCTP 和 dTTP）存在的条件下，引物始动新 DNA 链的合成，此链与靶 DNA 片段的单个 DNA 链互补。

审慎地选择引物方向，这样新链合成的方向从一个引物发生，朝向另一个引物的结合位点。结果，新合成链可以轮流充当新 DNA 合成的模板，引起产物呈指数增加的连锁反应（图 5.2）。

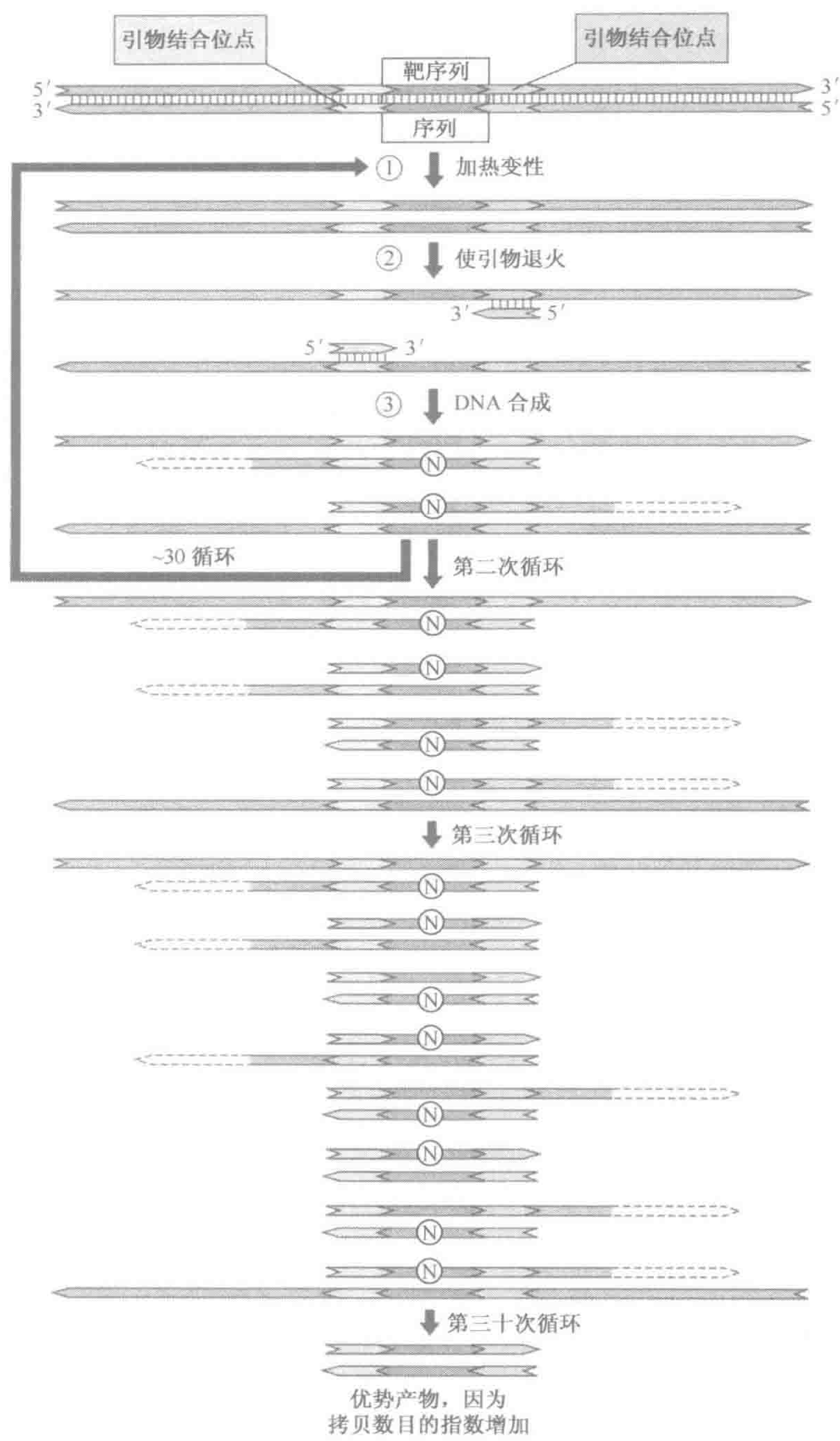


图 5.2　PCR 是一种利用限定的寡核苷酸引物体外扩增 DNA 序列的方法

识别将要被扩增的序列（**靶序列**，target sequence）后，设计寡核苷酸引物与位于反向 DNA 链和靶序列旁侧的 DNA 序列互补。PCR 由变性、引物的复性，然后是 DNA 合成的循环组成，循环中，引物被掺入新合成的 DNA 链。第一轮循环产生两条新 DNA 链（N），其 5′端固定在寡核苷酸引物的位置，但 3′端是可变的（用虚线表示）。第二轮循环后，四条新链由带有可变异 3′端的两个产物组成，就像第一轮循环一样，但是两条新链的固定长度（都携带 5′端和 3′端）由引物序列限定。第三轮循环后，八条新链中的六条具有预期的固定长度，30 次循环左右，此型产物大量增加（扩增）。

对热稳定酶的需求促使研究人员从天然栖息地温泉的微生物中提取 DNA 聚合酶。早期广泛应用的例子是由水生致热菌获得的 Taq 聚合酶，对热稳定性直至 94℃，最适工作温度是 80℃。然而，Taq 聚合酶缺乏相关的可以提供**校对功能**（proofreading function）的 3′→5′外切核酸酶活性，所以与细胞内发生的情况相比，拷贝错误（掺入错误碱基）频发。结果，一个具有 3′→5′ 校对功能的外切酶活性的替代酶，例如来自于热稳定性细菌的 Pfu 聚合酶目前被广泛使用（Cline *et al.*,1996）。

5.2.2 PCR 有两个主要限制：短片段和低产量

PCR 作为 DNA 克隆方法的明显缺点是扩增产物的大小范围。与细胞 DNA 克隆不同，克隆 DNA 序列大小的上限可达 2Mb，报道的 PCR 产物典型地是在 0～5kb 大小范围内，且经常处于此范围的低限。尽管 DNA 的小片段通常可以利用 PCR 很容易地被扩增，但是，随着目的产物长度的增加，更加难以获得有效的扩增。尽管如此，已经开发了**长范围 PCR**（long-range PCR）方法，产生长度达上万个碱基的产物，例如 λ 噬菌体的 42kb 产物（Cheng *et al.*,1994）。修饰条件经常与两个类型的对热稳定的 DNA 聚合酶混合物有关，努力提供 DNA 聚合酶和充当校对功能的 3′→5′ 外切核酸酶活性的最适水平。

对单一 PCR 反应中可以被克隆的物质的量也有限制，且多次重复相同的 PCR 反应获得大量的目的 DNA 是耗时的，并且费用昂贵。此外，PCR 产物可能不是进行某些后续研究的合适形式。因此，利用细胞克隆体系来克隆 PCR 产物是方便的，可以此获得大量的目的 DNA 进行各种分析。

利用各种质粒体系在细菌细胞中繁殖 PCR 克隆的 DNA。一旦克隆，可以利用合适的限制性核酸酶切掉插入序列，导入有特殊用途的其他质粒，表达 RNA 产物或大量蛋白质。包括 Taq 聚合酶在内的几种热稳定聚合酶有末端脱氧核苷酸转移酶活性，通过加入一单核苷酸选择性修饰 PCR 产生的片段，通常是腺核苷酸附加到被扩增的 DNA 片段的 3′端。

形成的突出末端使 PCR 产物难以克隆，通常使用的各种方法使克隆变得容易，包括使用在克隆位点多接头内具有突出 T 残基的载体（图 5.3），以及‘抛光’酶的使用，例如可以切掉突出的单核苷酸的 T4 聚合酶或 Pfu 聚合酶。或者，通过设计一在 5′端含有合适限制性位点约 10 个核苷酸长的序列可以修饰 PCR 引物。扩增期间，核苷酸延伸不与靶 DNA 进行碱基配对，但是后来可以用合适的限制性酶消化被扩增的产物，产生突出末端以便克隆导入合适的载体（节 5.5.3）。

5.2.3 PCR 的一般应用

由于其简单性，PCR 是应用广泛受欢迎的技术，主要有三个优点：

- ▶ 快速，易于操作；
- ▶ 非常灵敏，能够扩增微量靶 DNA——甚至单细胞 DNA（Li *et al.*,1988）。已经发现 PCR 广泛应用于诊断，遗传连锁分析（包括单精子分型）和法医学（个体遗留的痕量组织——例如一根毛发或丢弃的皮肤细胞——可以利用高度多态性 DNA 标记进

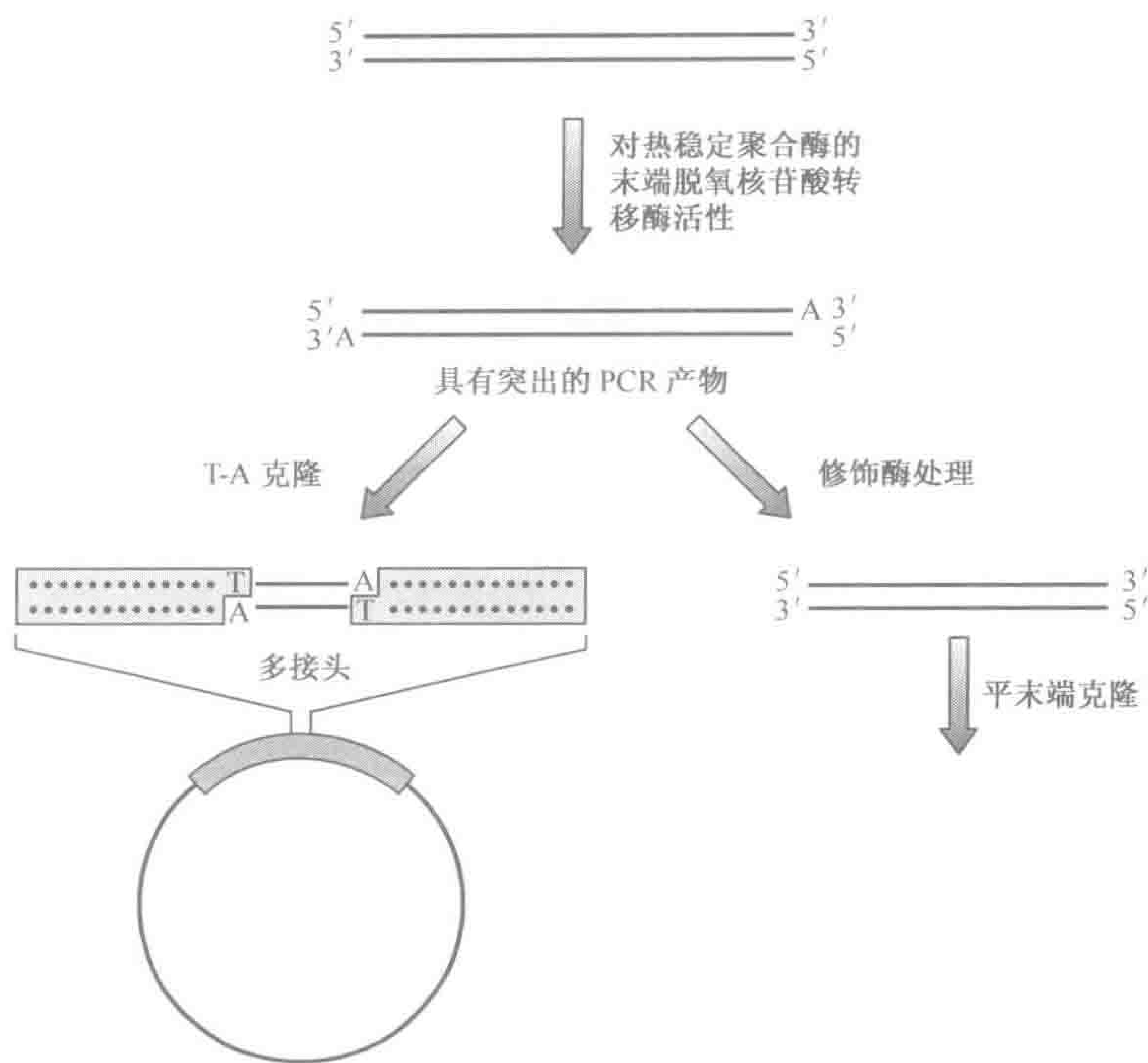


图 5.3　细菌细胞中 PCR 产物的克隆

PCR 产物经常在 3′端有突出的腺苷酸（见本文）。T-A 克隆体系包括携带互补的胸苷酸突出物的多接头体系，使克隆易于进行。另一种选择方法是用合适的"抛光"酶修饰腺苷酸突出物，保留平-末端片段。

行 PCR 分型，进而识别个体起源）。

▶ 能力强大，经常能够扩增严重降解的组织或细胞 DNA，或包埋在某些介质中用常规方法难以分离的 DNA。PCR 可以从小量降解的 DNA 中扩增短序列，这种 DNA 可以是从考古学或有历史的腐烂的组织中提取的，并且能够在石蜡固定组织成功地进行 PCR 扩增（这些组织对分子病理学，以及在某些情况下的遗传连锁研究具有重要作用）。

PCR 的许多应用，以及对优化效率和特异性的需求已经促使各种广泛的 PCR 方法应运而生（框 5.1）。由于引物特异性的至关重要性，经常使用各种修饰减少非特异引物结合的机会（例如，使用热启动 PCR，巢式引物，降落 PCR；框 5.1）。

对在结合引物的 3′端正确的碱基配对的严格依赖，已经发展了可以区别两个等位基因之间仅单个核苷酸差异的方法（**等位基因特异性 PCR**，allele-specific PCR）。在流行的**难以扩增的突变系统**（amplification refractory mutation system，ARMS）方法中，设计引物的 3′ 端核苷酸与区别两个等位基因的变异核苷酸进行碱基配对，引物的其余序列与最邻近变异核苷酸的序列互补。在合适的实验条件下，扩增不会发生在 3′ 端核苷酸碱基配对不完全的地方，因此两个等位基因得以区别（图 5.4）。

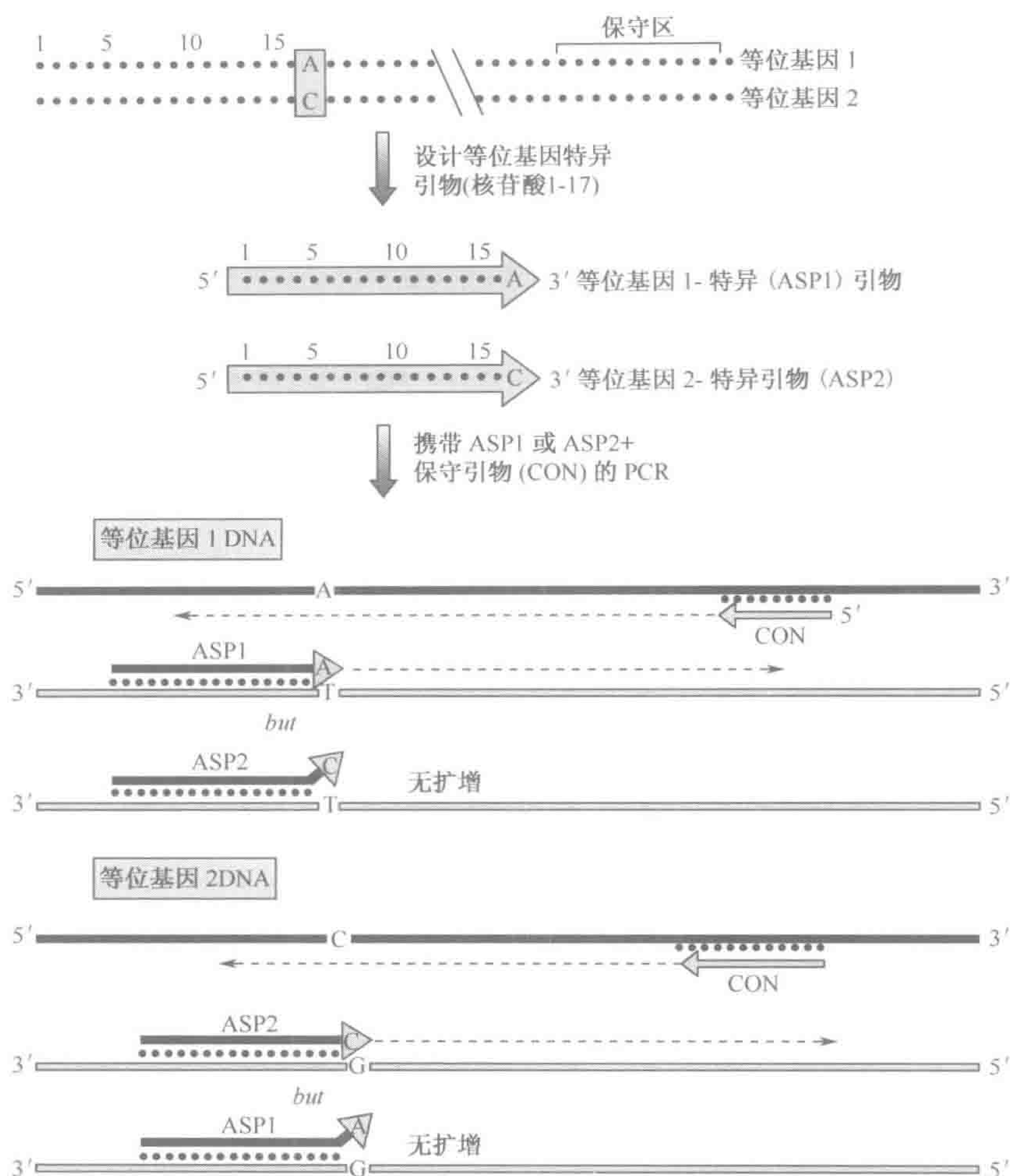


图 5.4 使用 ARMS 体系的等位基因特异 PCR 取决于引物的 3′端核苷酸的完全碱基配对

设计等位基因特异寡核苷酸引物 ASP1 和 ASP2 与变异核苷酸之前片段上的两个等位基因序列相同，延伸并终止在变异核苷酸自身。ASP1 将与等位基因 1 序列的互补链完全结合，可以进行带有保守引物的扩增。然而，ASP2 引物 3′端的 C 与等位基因 1 序列的 T 错配，使扩增不能进行。与 ASP1 不一样，类似的 ASP2 可以与等位基因 2 完全结合并始动扩增。

5.2.4 设计某些 PCR 反应产生多重扩增产物并扩增先前未鉴定的序列

标准 PCR/RT-PCR 反应是基于对特异引物结合的需求，目的是选择性扩增已知的目的靶序列。然而，有时想要设计 PCR 反应扩增了解有限的或事先未提供序列信息的 DNA 序列。

利用 DOP-PCR 扩增 DNA 家族新成员

如果先前未鉴定的 DNA 序列是一个基因或重复 DNA 家族的成员，但至少其成员之一先前已鉴定过，有时可以利用 PCR 克隆先前未鉴定的 DNA 序列。例如，在注意到前几个 *Wnt* 基因产物在特殊的保守结构域内氨基酸序列十分相似后，哺乳动物 *Wnt* 基因家族的许多新成员首次被分离。由于在每种情况下，不同寡核苷酸的混合物代表各

种氨基酸序列改变，这就允许基于**变性寡核苷酸**（degenerate oligonucleotide）设计该序列的引物。**变性寡核苷酸引导的PCR**（DOP-PCR）可以允许同时扩增各种不同的但密切相关的基因，包括新基因，然后通过细胞DNA克隆被分离和纯化。

无区别扩增

如果DNA资源宝贵且数量十分有限，有可能通过将双链**接头寡核苷酸**（linker oligonucleotide）共价连接到起始群体的所有DNA序列末端，利用PCR扩增所有DNA。单个合成两个寡脱氧核苷酸来制备接头寡核苷酸，设计两个寡脱氧核苷酸在序列上互补且能够碱基配对形成携带突出末端的双链DNA序列。用限制性核酸酶消化将要被扩增的DNA，产生相似的突出末端，这样接头能共价连接（连接）。接头特异引物允许扩增两端携带接头的靶DNA分子。结果，起始序列中的所有DNA序列可以同时扩增，就基因组DNA而言，可以扩增**全基因组**（whole genome amplification）。另一种方法是利用广泛变性的寡核苷酸作为引物，这样引物可以与许多结合位点结合。

利用已知的序列扩增邻近未知的DNA序列

从已知起始DNA序列到邻近未知的序列，无论是基因组DNA还是cDNA，人们已经设计进行各种巧妙的修饰。例子包括**锚定PCR**（anchored PCR），**反向PCR**（inverse PCR），**快速PCR**（race-PCR）（框5.1）。

5.3　细胞DNA克隆原理

5.3.1　细胞DNA克隆的概况

细胞DNA克隆最早是在20世纪70年代早期发展起来的。限制性内切核酸酶Ⅱ型的发现使细胞克隆成为可能，它是一细菌酶，能够在含有小的特异识别序列——通常4～8bp长的所有位点——切割DNA。正常情况下，这些酶通过选择性切割外源性DNA保护细菌免受噬菌体的侵袭（见框5.2和下一节）。它对分子遗传学的最大贡献是提供一种将DNA切割成限定的片段，使之能够易于与其他被切割的类似DNA片段连接的方法。

框5.2　限制性内切核酸酶和修饰-限制体系

从一特殊菌株释放的噬菌体可以感染除了不同菌株外的相同菌株的其他细菌。这是因为就像可以感染的菌株DNA一样，噬菌体DNA有相同的**修饰**（modification）模式；噬菌体被“**限制**”（restricted）到那个菌株。限制不是绝对的：某些噬菌体可以逃离限制并获得新宿主的修饰模式。众所周知，修饰-限制体系的基础与两种类型的酶活性有关：

- 序列特异的**DNA甲基化酶**（DNA methylase）活性奠定修饰模式的基础；
- 序列特异的**限制性内切核酸酶**（restriction endonuclease）活性奠定限制现象的基础：此酶切割与宿主细胞DNA甲基化模式不同的噬菌体DNA。

与相应的限制性核酸酶活性一样，菌株具有相同序列特异性的甲基化酶活性。结果，细胞的限制性内切核酸酶不会切割适当甲基化的宿主细胞DNA，但可以切割外来的噬菌体DNA，前提是没有发生适当的甲基化。

注：某些质粒和噬菌体具有修饰和限制体系基因，可以赋予宿主细胞这种专一性。

细胞 DNA 克隆的精髓是利用限制性核酸酶在起始 DNA 群体将 DNA 分子（**靶 DNA**，target DNA）切割成可控制大小的片段，然后与**复制子**（replicon）（能够进行独立 DNA 复制的任何序列）连接并将合成的杂交分子（**DNA 重组**，recombinant DNA）导入合适的宿主细胞，然后经细胞分裂而增殖。因为复制子可以在细胞中复制（经常达高拷贝数目），所以连接的靶 DNA 也是如此，结果导致细胞 DNA 的扩增。

细胞 DNA 克隆有四个主要步骤（图 5.5）：

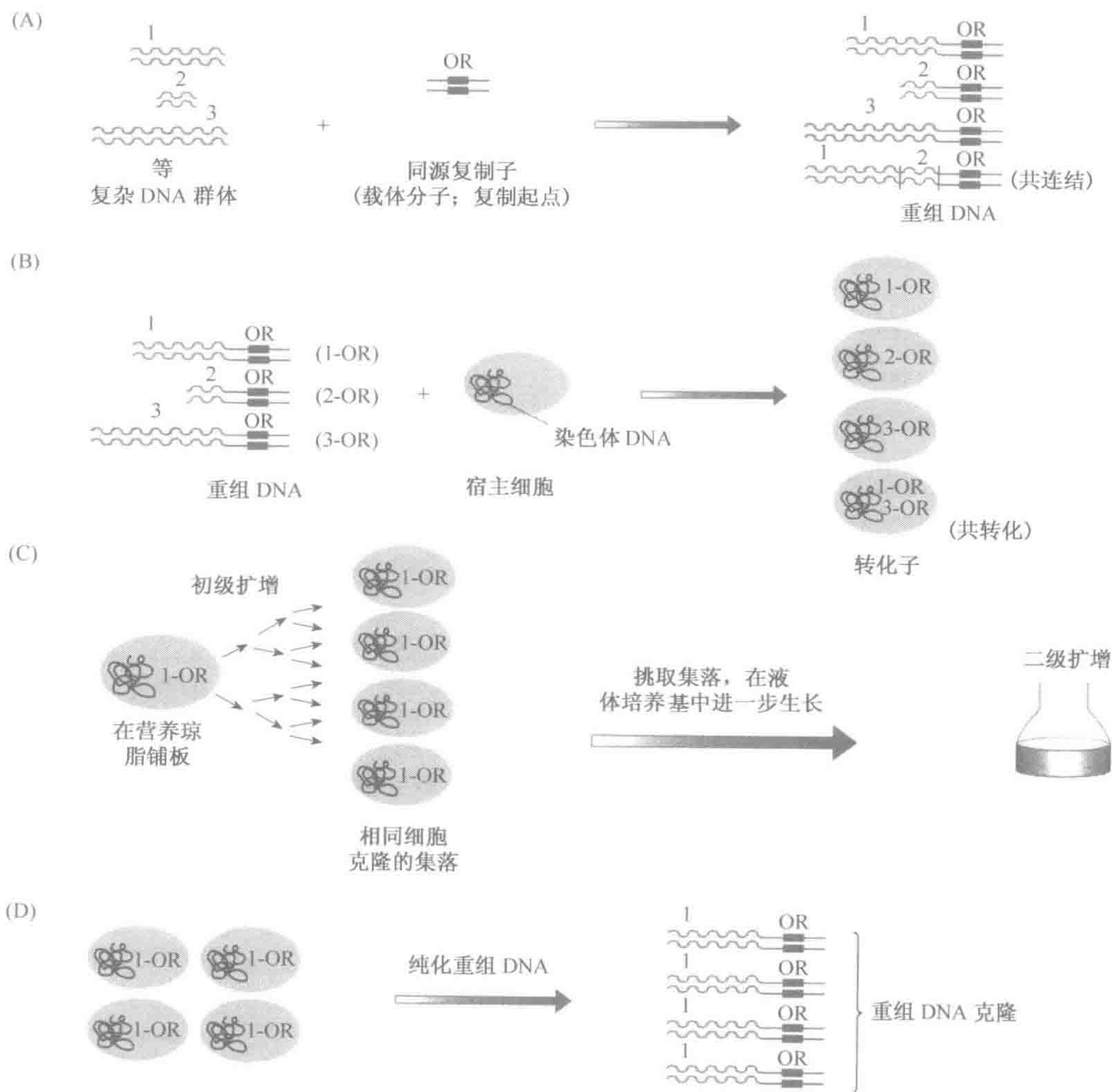


图 5.5　细胞 DNA 克隆的四个主要步骤

(A) **DNA 重组形成。**注意，除了简单的载体-靶连接产物以外，凭借两个不相关的靶 DNA 序列可能被连接成一个产物（例如下面例子的序列 1 和序列 2），共连接事件可能发生。OR，复制起点。(B) **转化**。这是 DNA 克隆的关键步骤，因为正常情况下，细胞仅接纳一个外源 DNA 分子。然而值得注意的是，偶尔观察到共转化事件，如下面被阐述的细胞通过两个不同的 DNA 分子（含有序列 1 的重组分子和含有序列 3 的重组分子）被转化。(C) **扩增产生许多细胞克隆。**这是另一个关键步骤。转化细胞铺板后，可以在平皿上分离单个克隆集落，然后，挑取单个克隆进行第二次扩增步骤，以确保克隆纯一性。(D) **DNA 重组**克隆的分离。

▶ **重组DNA分子的构建**　通常利用限制性核酸酶切割产生大小合适的DNA片段，随后将靶DNA片段共价连接（连接，ligation）到单一类型**复制子**（replicon）分子上。利用DNA连接酶将靶DNA片段连接到复制子分子之前，通过保证用产生同型末端的限制性核酸酶切割靶DNA和复制子分子，则DNA重组的构建易于进行；

▶ **转化**　DNA重组分子导入宿主细胞（一般是细菌或酵母细胞），其中选择的复制子可以进行独立于宿主细胞染色体的DNA复制。细胞克隆使用的复制子经常被称为**载体分子**（vector molecule），因为它们有助于运输目的靶分子进入细胞，然后在细胞内复制；

▶ **细胞克隆的选择性繁殖**涉及两个阶段。最初，转化细胞被平铺在琼脂表面，便于促进分离很好的细胞集落生长。单细胞集落的所有细胞完全相同（因为他们来自同一个细胞），被称为**细胞克隆**（cell clone）。随后，可以从平皿中挑取单个集落，在液体培养基中培养使细胞进入生长的第二阶段；

▶ **DNA重组克隆的分离**是通过收获增殖的细胞培养基，选择性分离DNA重组。

5.3.2　限制性内切核酸酶能使靶DNA切割成易操纵的片段并可与类似切割的载体分子连接

限制性核酸酶Ⅱ型

正常情况下，绝大多数限制性内切核酸酶Ⅱ型的识别序列是**回文结构**（palindrome）（当按5′→3′方向阅读时，两条链的碱基序列是相同的。是双重对称轴的结果）。取决于限制性核酸酶产生的切割位点的位置，形成的**限制片段**（restriction fragment）可以有：

▶ **平末端**（blunt end）（切割点准确发生在对称轴上）；

▶ **突出末端**（overhanging end）[切割点不在对称轴上，所以，形成的限制片段具有所谓的5′**突出末端**或3′**突出末端**（表格5.1）]。

注：每个片段的两个突出末端在碱基序列上互补，且有相互联系的倾向（或具有任何其他类似的互补的突出末端）。结果易于黏附到同一类型的其他末端，此型的突出末端经常被描述为黏性末端。

限制性内切核酸酶Ⅱ型为DNA克隆和DNA分析提供两个优点：

▶ **一套限定的起始DNA片段**　当从组织和培养细胞分离DNA时，巨大DNA分子的打开和不可避免的机械撕裂产生随机的DNA片段长度不同的不均一性群体。利用限制性内切核酸酶，有可能将这种巨大的不均一的破碎DNA片段转化成限定长度的**限制片段**（restriction fragment）；

▶ **人工DNA重组分子的方法**　利用DNA连接酶，具有相同黏性末端的限制片段可以易于连接在一起。因此，要想构建DNA重组，利用同型限制性核酸酶能够切割复制子（**载体**，vector）分子和靶DNA，或利用限制性核酸酶产生同类型的黏性末端。

有相同类型突出末端的限制性片段末端能够以各种不同方式连接，或分子内（**环化**，cyclization），或分子间形成线性**多连体**（concatemer）或环状化合物分子。在高DNA浓度时，分子间反应大多数易于发生。然而，DNA浓度极低时，不同分子的单个

末端很少有机会相互接触，有利于分子内环化。尽管也有可能出现载体环化，载体一载体多连体以及靶 DNA—靶 DNA 连接是可能的（图 5.6），但通常设计连接反应促进 DNA 重组形成（通过将靶 DNA 连接到载体 DNA）。为了达到此目的，载体分子经常被处理，以预防或最小化进行环化的能力。

表 5.1　常用的限制性内切核酸酶的例子（并参见表 6.3 罕见切割）

酶	来源	切割序列 N=A、C、G 或 T	具有以下末端的限制片段
产生平末端			
AluI	藤黄节杆菌	A↓GCT	5′CT——AG3′
	(Arthrobacter luteus)	T↑CGA	3′GA——TC5′
产生 5′黏末端			
EcoRI	大肠杆菌 R 因子	↓GAATTC	5′AATTC——G3′
	(Escherichia coli R factor)	CTTA↑AG	3′　G——CTTAA5′
产生 3′黏末端			
PstI	斯氏普罗威登斯菌	CTGC↓AG	5′G——C　3′
	(Providencia stuartii)	↑GACGTC	3′　C——G5′
识别非回文序列			
Mnl I	奥斯陆莫拉菌	CCTCNNNNNNN↓N	5′——CCTCNNNNNNN3′
	(Moraxella nonliquefaciens)	GGAGNNNNNN↑NN	3′N—GGAGNNNNNN　5′
识别对分识别序列			
BstXI		CCANNNN↓NNTGG	5′NTGG—CCANNNNN3′
		GGT↑NNNNNNACC	3′　NNNNNACC—GGTN5′

注：正常情况下，名字来源于种的第一个字母和种属的头两个字母。例如 Psi 是第一个从 *Providencia stuart* Ⅱ 提取的限制性核酸酶——见限制性核酸酶的 REBASE 数据库http://rebase.neb.com/rebase/rebase.html。

在细菌细胞中克隆的简单载体

在细胞 DNA 克隆期间，靶 DNA 片段必须能够在细胞内复制。因为它们缺少功能性的复制起点，所以他们需要被连接到可以在宿主细胞内独立于宿主细胞染色体复制的复制子（载体）上。载体可以有一个起源于天然染色体外复制子的复制起点，或在某些情况下，起源于染色体复制子的复制起点（正如在酵母人工染色体的情况下，见节 5.4.4）。

最频繁使用的宿主细胞是修饰的细菌或真菌宿主细胞。细菌细胞宿主因其快速的细胞分裂能力而尤为广泛应用。它们具有携带单一复制起点的单环双链染色体。宿主染色体复制随后引起细胞分裂，这样两个子代细胞中的每一个就像他们的双亲一样含有单染色体（例如，拷贝数维持每个细胞一个拷贝）。然而，染色体外复制子的复制并不局限于这种方式：在细胞周期期间，许多这样的复制子经历几个复制循环并能够达到高拷贝数。结果，大量靶 DNA 通过与复制子共复制可以制备。染色体外复制子通常分为两类：

▶ **质粒**（plasmid）小的环状双链 DNA 分子，含有很少的基因。质粒存在于细胞内，

伴随宿主细胞分裂，垂直分配给子细胞。但在细菌接合事件中，可以横向传递给邻近细胞。天然例子包括携带性因子（F）的质粒以及那些携带耐药基因的质粒。

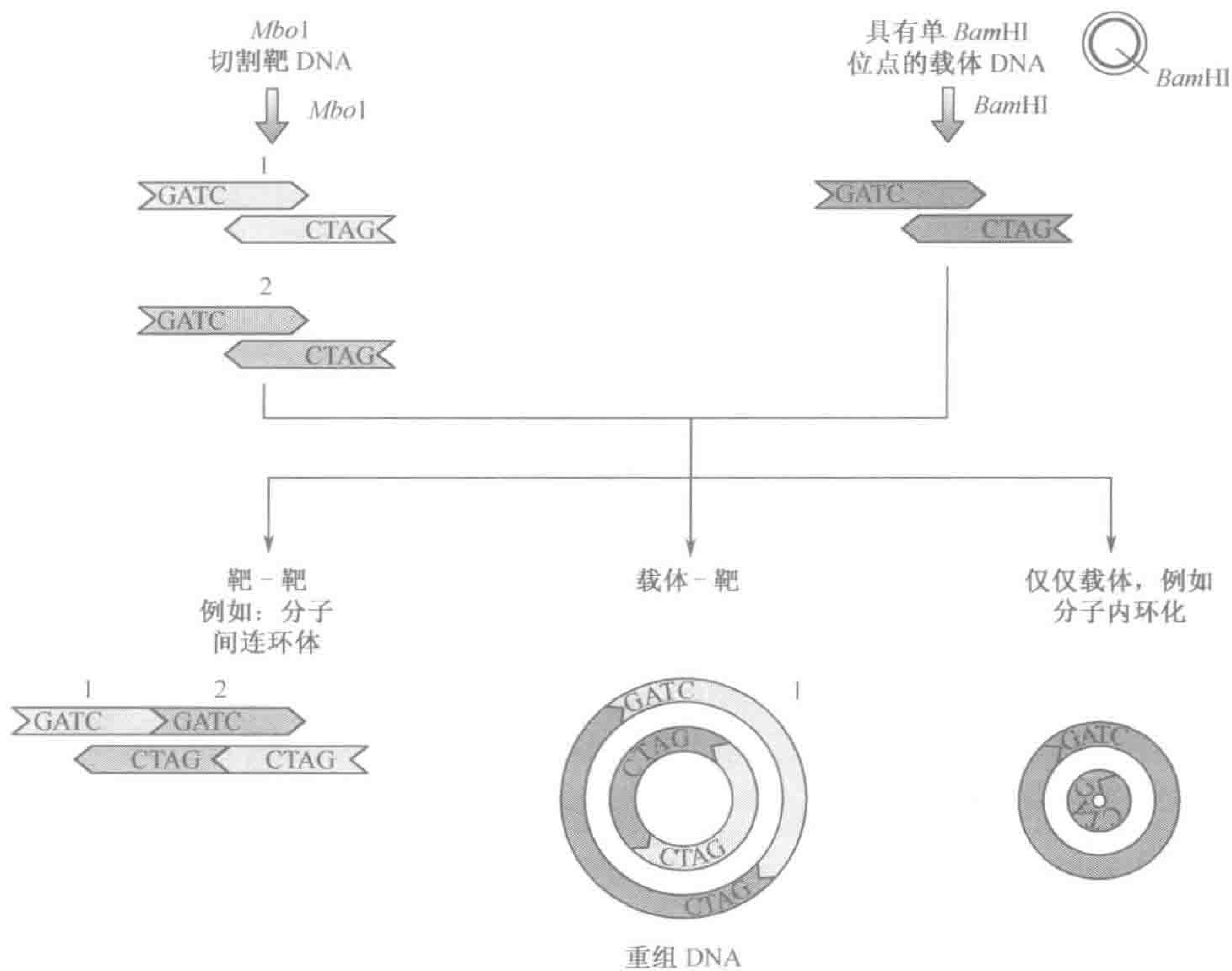


图 5.6　黏附末端可以在分子内和分子间连接

注：仅显示某些可能的结果。例如，载体分子也可以形成分子间多连体，多聚体可以进行环化，且共连接事件涉及两个不同的靶序列，后者与载体分子一起被包括在同一 DNA 重组分子内（图 5.5A）。当 DNA 处于低浓度时，个体分子环化的倾向更显著，且携带互补黏性末端的不同分子间的碰撞机会减少。

▶ **噬菌体**（bacteriophage）——感染细菌细胞的病毒。含有 DNA 的噬菌体经常具有环状或线状双链 DNA 的基因组。与质粒不一样，它们可以存在于细胞外。成熟病毒颗粒（**病毒粒子**，virion）有自己的基因组，包装在蛋白外壳内，以致易于吸收并进入新的宿主细胞。

5.3.3　将 DNA 重组导入受体细胞为分离复杂的起始 DNA 群体提供了一种方法

质粒细胞膜选择性渗透，在正常情况下不允许大分子如长的 DNA 片段透过。然而，细胞能够用某些方法处理（例如，暴露给某些高离子强度盐，短暂的电子休克等），这样质膜的渗透性特性将会改变。结果，小部分细胞处于**易感状态**（competent），这意味着细胞能够接纳来自细胞外环境的外源 DNA。

只有小部分细胞会接纳外源 DNA（DNA **转化**，transformation）。然而，那些真正接纳外源 DNA 的细胞经常只摄取单个分子（然而，单个分子随后可以在细胞内复制许多次）。这就是细胞 DNA 克隆中主要分离步骤的基础：转化细胞的群体可能被认为是

分类机构，其中，复杂 DNA 片段混合物借助个体 DNA 分子存入受体细胞而被分类（图 5.7）。

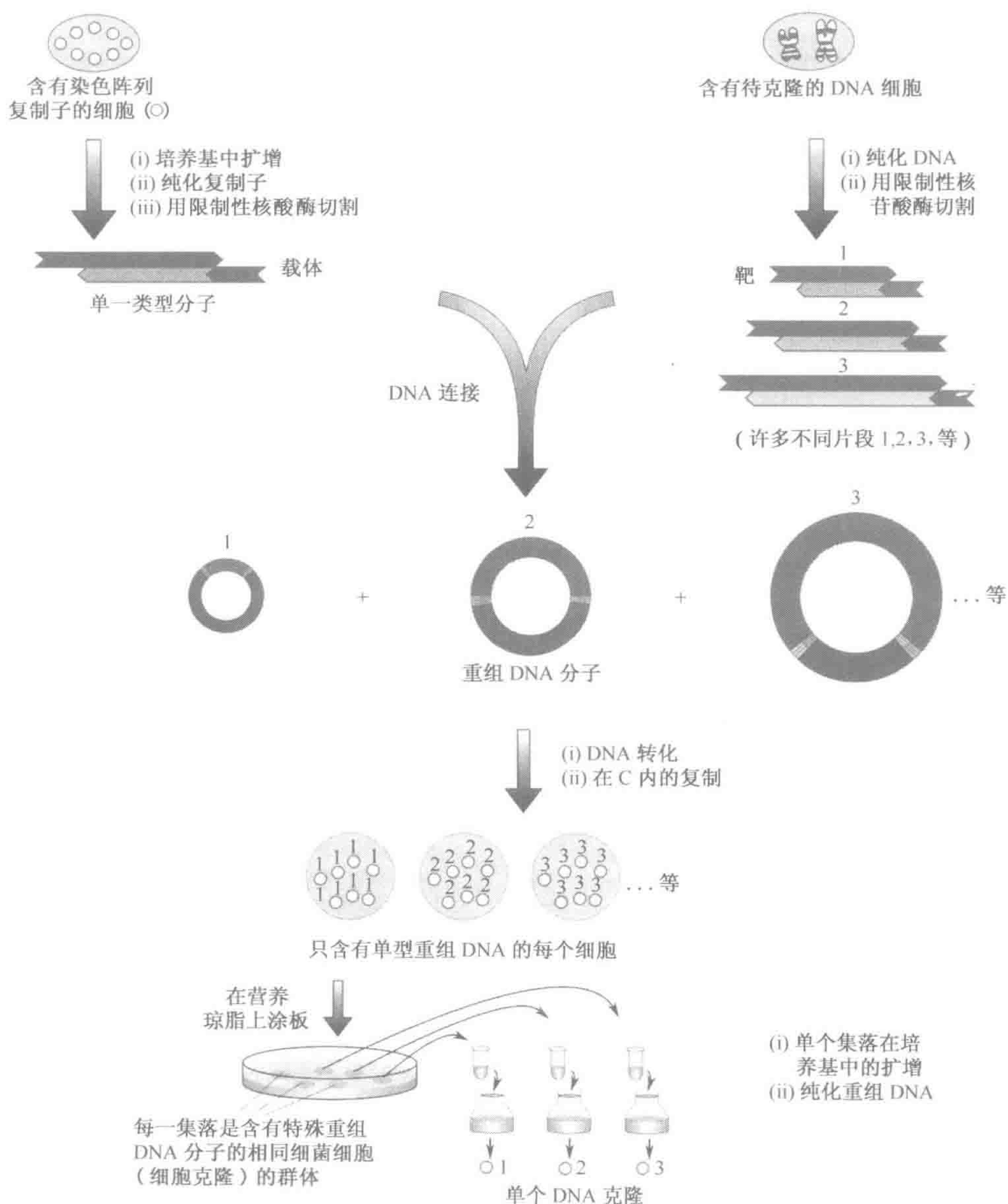


图 5.7　细菌细胞的 DNA 克隆

举例阐述基因组 DNA 克隆，但同样应用于克隆 cDNA 。

因为环状 DNA（甚至缺口环状 DNA）转化比线状 DNA 分子的转化更有效率，所以大多数转化子含有环状产物而不是线状 DNA 重组多连体，并且如果试图抑制载体环化（例如，通过去磷酸化），大多数转化细胞株将含有 DNA 重组。然而，注意**共转化**(cotransformation)事件（在一个细胞克隆内导入一种以上类型的 DNA 分子的事件，

见图 5.5B）在某些克隆体系可能比较常见。

允许转化细胞繁殖。就利用质粒载体和细菌宿主细胞进行克隆而言，含有转化细胞的溶液简单涂在培养皿的营养琼脂表面（**涂板**，plating out）。这通常导致由**细胞克隆**（cell clone）（单一始祖细胞的相同子代）组成的**细菌集落**（bacterial colony）的形成。挑取单个集落放入管内，随后在液体培养基内生长，细胞可以进行第二次扩增，提供高产量的细胞克隆，所有细胞均与始祖单细胞相同（图 5.7）。如果起始细胞含有附着于复制子的单一类型外源 DNA 片段，那么子代也是如此，结果导致特异外源 DNA 片段大量扩增。然后，代表起源于单细胞的细胞克隆的扩增培养基可以进行以恢复 DNA 重组。

为了恢复从裂解细胞选择的 DNA 重组，可以利用宿主细胞 DNA 和 DNA 重组之间的有形区别。就细菌细胞而言，双链细菌染色体像任何含有导入外源 DNA 的质粒一样，是环状的，但片段非常大。结果，在细胞裂解及随后的 DNA 提取时，易于形成缺口和撕裂，产生具有游离末端的线状 DNA 片段。分离的 DNA 变性后，例如通过碱性处理，线状宿主细胞 DNA 容易变性，**但密闭的共价的环状**（covalently closed circular，CCC）质粒 DNA 不能分离，当再次复性时，两条链又重新连接形成天然的**超螺旋**（superhelical）分子或所谓的**超螺旋 DNA**（supercoiled DNA）（图 5.8）。变性的宿主细胞 DNA 从溶液中析出，留下 CCC 质粒 DNA。

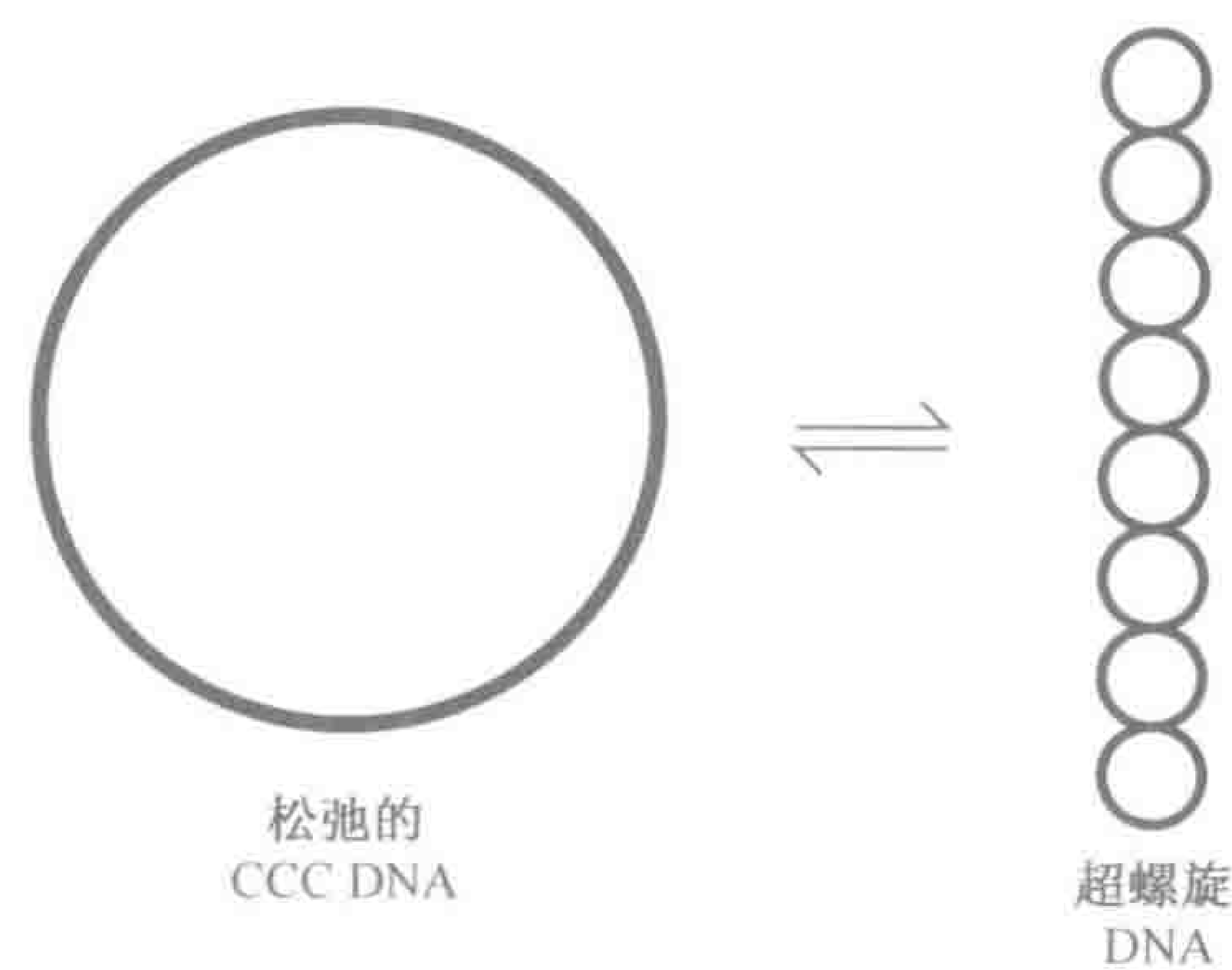


图 5.8　密闭共价环状（CCC）DNA 和 DNA 超螺旋

图表左侧显示 CCC DNA（没有显示双螺旋结构）。除非两条 DNA 链中的一条有缺口，否则双螺旋的缠绕不能解开，紧张诱导引起超螺旋结构自发形成显示在右侧。然而，DNA 链的缺口解除了游离端旋转引起的张力。

如果需要的话，有可能通过**平衡密度梯度离心**（**等密度离心**，isopycnic centrifugation）进一步纯化：通过离心在含有溴化乙啶（EtBr）的氯化铯溶液中平衡部分纯化 DNA。EtBr 通过碱基对之间的相互螯合结合 DNA，由此引起 DNA 螺旋解旋。与染色体 DNA 不一样，CCC 质粒 DNA 没有游离末端，且只能解旋到有限的程度，这限制了其可以结合 EtBr 的量。当 EtBr- DNA 复合物含有较少 EtBr 时，复合物密度较大。所以，CCC 质粒 DNA 比染色体 DNA 或开放的质粒环在氯化铯梯度中结合的位置更低，能够从宿主细胞分离 DNA 重组。形成的 DNA 重组分子通常会彼此相同（代表单一靶

DNA 片段)，被称为 **DNA 克隆** (DNA clone)。

5.3.4 DNA 文库是代表复杂起始 DNA 群体的全套 DNA 克隆

第一次尝试在细菌细胞中克隆人类 DNA 片段主要集中在高度富含在特殊的起始 DNA 群体中的靶序列。例如，人类的有核细胞含有许多相同的 DNA 序列，但 mRNA 群体可能完全不同。尽管每个细胞的 mRNA 群体是复杂的，但是某些细胞专一合成特殊类型的蛋白质，所以他们有几个居优势的 mRNA 种类 (例如，在红细胞中制备的许多 mRNA 是由 α 和 β 珠蛋白 mRNA 组成)。**反转录酶** (reverse transcriptase) (RT; RNA-依赖 DNA 聚合酶) 可以用来制备碱基序列与 mRNA 互补的 cDNA 拷贝。因此，来自红细胞的 cDNA 很大程度上富含于珠蛋白 cDNA，易于分离。

现代 DNA 克隆方法为从极其复杂的起始 DNA 群体 (例如人类总基因组 DNA) 中制备全套 DNA 克隆 (**DNA 文库**) (DNA library) 提供了可能性。这种方法使在起始群体中非常罕见的 DNA 序列可以在 DNA 克隆文库中体现出来。因此，通过选择合适的宿主细胞集落并对其进行扩增，可以单独分离罕见 DNA 序列。这种方法的两个基本类型已经被广泛采用，取决于起始 DNA 的性质：基因组 DNA 文库和 cDNA 文库。

据说新构建的文库是非扩增的，尽管这是一个使人误解的术语，因为起始转化细胞扩增形成分离的细胞集落。细胞集落的形成经常发生在膜表面，这些膜被平铺在无菌培养皿的营养琼脂表面。通过将复制拷贝涂在大小相似事先铺在营养琼脂表面的膜上并呈集落生长，以此来制备文库拷贝。最近，单个挑取的细胞集落已经被点在合适膜的格阵里或微量滴定平皿的孔内进行斑点杂交，这样它们可以在使细胞稳定的培养基诸如甘油中于－70℃长期贮存。

对于多重分布，需要**扩增的文库** (amplified library)。将典型的原始滤膜上的细胞洗掉，放入培养基中，稀释并借助甘油或一些可选择的稳定剂使其稳定。然后，单个分装在后期被涂板，重新制备文库。然而，这个附加的扩增步骤可能导致细胞克隆的原来特征的破坏，因为在扩增阶段，不同集落的生长率可能不同。

基因组 DNA 文库

就复杂的真核生物而言，例如哺乳动物，所有的有核细胞实质上有相同的 DNA 含量，且很方便从易于获得的细胞诸如白细胞制备基因组文库。起始材料是基因组 DNA，可以用某种方法将其切割成片段，通常是用限制性内切核酸酶消化。

典型地，基因组 DNA 用 4bp 切割工具消化，诸如识别序列 GATC 的 *Mbo*I。人类基因组 DNA 平均约每 280bp 会出现此序列，所以很少有 DNA 序列缺乏该酶的识别位点。用 *Mbo*I 完全消化起始 DNA 会产生非常小的片段。取而代之，进行**部分限制性消化** (partial restriction digestion) (低的酶浓度，孵育时间短，等等)，所以断裂只发生在少数潜在的限制位点。

部分限制性消化不仅用于产生克隆的目的大片段，但重要的是，也产生随机 DNA 片段。因此，针对某一特异序列位点，相同的起始 DNA 序列的拷贝数不同，切割模式会有所不同 (图 5.9)。这样的随机片段确保文库尽可能含有和起始 DNA 同样多的拷贝。另外，随机片段的优点是它产生重叠插入片段的克隆。结果，每个克隆插入片段鉴

定后，通过识别那些携带与原来的克隆插入片段有某些相似性的插入片段，努力获得来自相同的一般区域的克隆（框 8.5）。

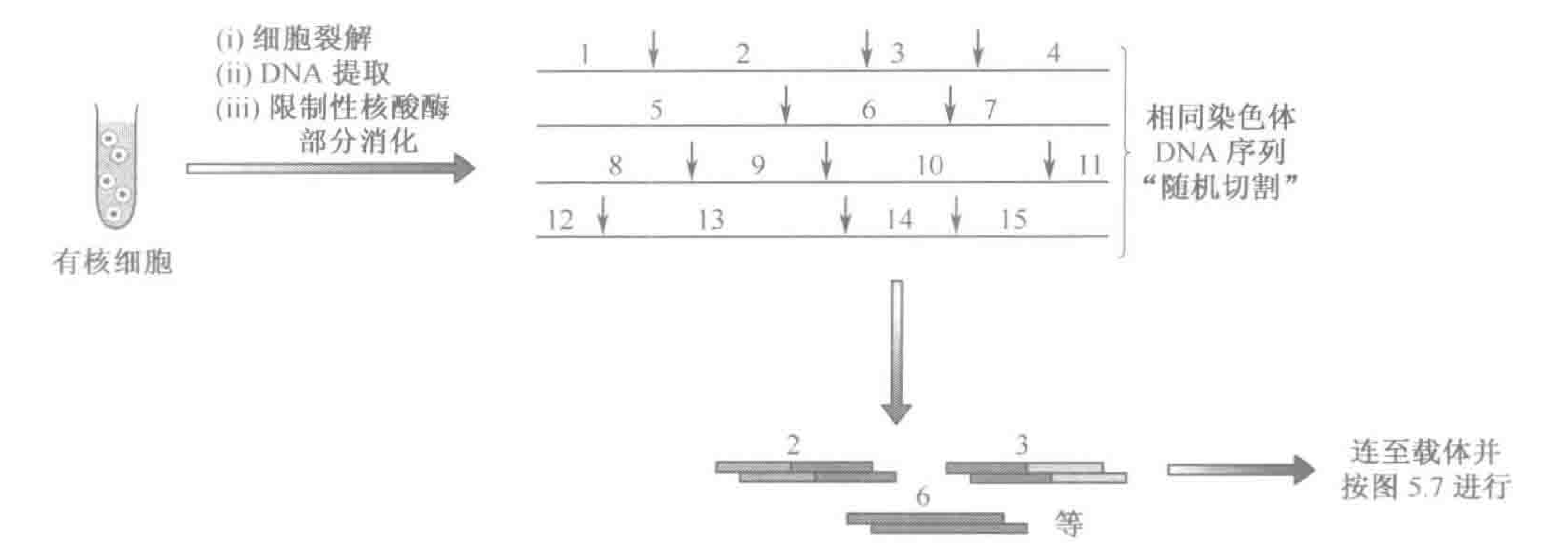


图 5.9　制备基因组 DNA 文库

个体的所有有核细胞有相同的基因组含量，所以任何易于得到的细胞（如白细胞）都可以作为原材料。因为 DNA 是从携带相同 DNA 分子的无数细胞中提取的，被分离的 DNA 将含有大量相同 DNA 序列。然而，限制性内切核酸酶的部分消化仅在可用的含有限制位点的小亚群切割 DNA，且个体分子间模式不同，结果几乎是随机切割。这样会产生一系列限制片段，如果来自相同位点，这些片段可以共享某些共有 DNA 序列（例如，片段 6 与片段 2 和片段 3 部分重叠，如图所示，片段 9 和 10，13，14 也是如此）。

一个基因组 DNA 文库的**复杂性**（complexity）（独立 DNA 克隆的数目）可以定义为**基因组当量**（genome equivalent，GE）一词。当独立 DNA 克隆数目＝基因组大小/平均插入大小时，基因组当量 1 称为**一倍文库**（one-fold library）。例如，对于携带平均插入大小 40kb 的人类基因组 DNA 文库，1 GE＝3000Mb/40kb＝75 000 个独立克隆。一个克隆诸如有 300 000 个克隆的文库有时称为四倍文库，因为它有 4GE。然而，由于样品多样化，GE 数目必须比 1 大很多时，任何特殊序列才有很高的机会包含在文库内。结果，通常试图制备复杂的（>4GE）文库。

cDNA 文库

因为在不同的细胞和不同的发育阶段，基因表达可能不同，制备 cDNA 文库的起始原料通常是来自某一特殊组织或胚胎发生的特殊发育阶段的总 RNA。由于绝大多数 mRNA 是多聚腺苷酸化的，所以 poly(A)$^+$ mRNA 通过与互补的 oligo(dT)或poly(U)序列特异结合，进而连接到固体琼脂糖或纤维素基质而被选择。然后，分离的 poly(A)$^+$ mRNA 利用反转录酶可以转化成双链 cDNA 拷贝。为帮助克隆化，含有合适限制位点的双链**寡核苷酸接头**（oligonucleotide linker）（**适配子**，adaptor）被连接到 cDNA的每一端（图 5.10）。

5.3.5　经常通过标记基因的插入失活完成重组筛查

对细胞 DNA 克隆体系的基本需求是检测含有合适载体分子的细胞以及其中含有 DNA 重组的亚群的一种方法。广泛筛查重组体对于筛查没有很好定性的 DNA 群体制备的 DNA 文库是有帮助的。然而，直接筛查可以特异检测某些事先已知序列或与已知

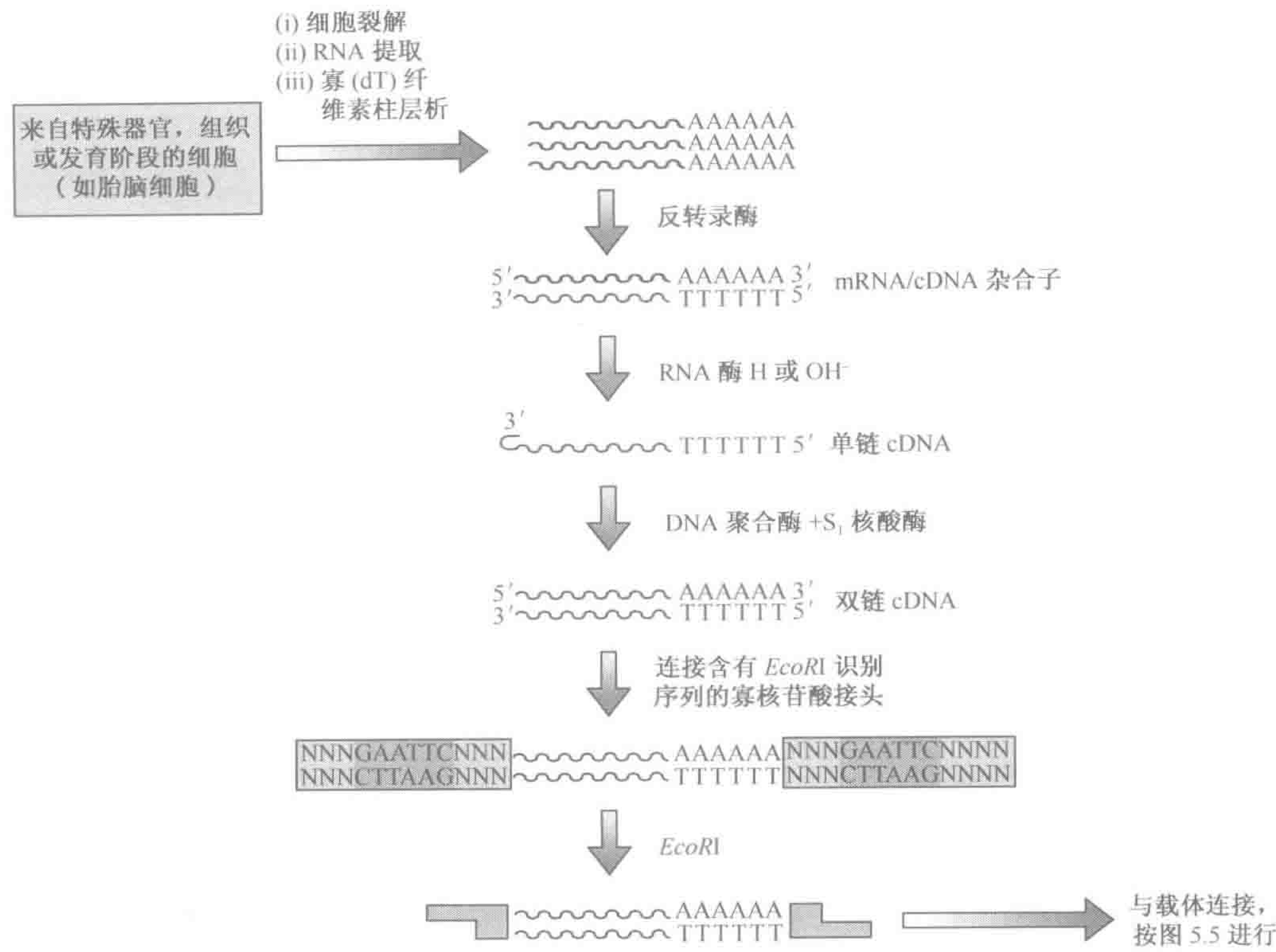


图 5.10　制备 cDNA 文库

反转录酶步骤经常使用 oligo（dT）引物指导 cDNA 链合成。最近，已经使用随机寡核苷酸引物替代，提供了更正常的序列代表。RNA 酶 H 会特异消化 RNA-DNA 杂交分子中与 DNA 结合的 RNA。结果，单链 cDNA 的 3′端趋于环袢形成短发夹。这可以用来借助 DNA 聚合酶指导第二条链合成，结果与两条链连接的短袢可被 S1 核酸酶切掉，S1 核酸酶特异地切割单链 DNA 区域。

序列密切相关序列的存在日益增加。

通过载体分子筛查转化细胞

识别含有载体分子的细胞需要设计或选择携带合适的标记基因的载体分子，标记基因的表达提供识别细胞的手段。两个普遍使用的标记基因体系是基于：

- **抗生素耐药基因**　选择使用的宿主细胞株对特殊抗生素敏感，经常是氨苄青霉素，四环素或氯霉素。设计相应的载体含有对抗生素耐受的基因。转化后，细胞涂在含有抗生素的琼脂上，通过载体筛选转化的细胞。
- **β 半乳糖苷酶基因互补性**　宿主细胞是含有 β 半乳糖苷酶基因片段的突变体，但不能产生有任何功能的 β 半乳糖苷酶。设计载体含有不同的 β 半乳糖苷酶基因片段。转化后，发生功能互补：宿主细胞和载体编码的 β 半乳糖苷酶片段能够结合，产生活化酶。具有功能的 β 半乳糖苷酶活性通过将无色底物——Xgal（5-bromo，4-chloro，3-indolyl β-D-galactopyranoside）——转化成蓝色产物而被检测。

一般性的重组体筛查

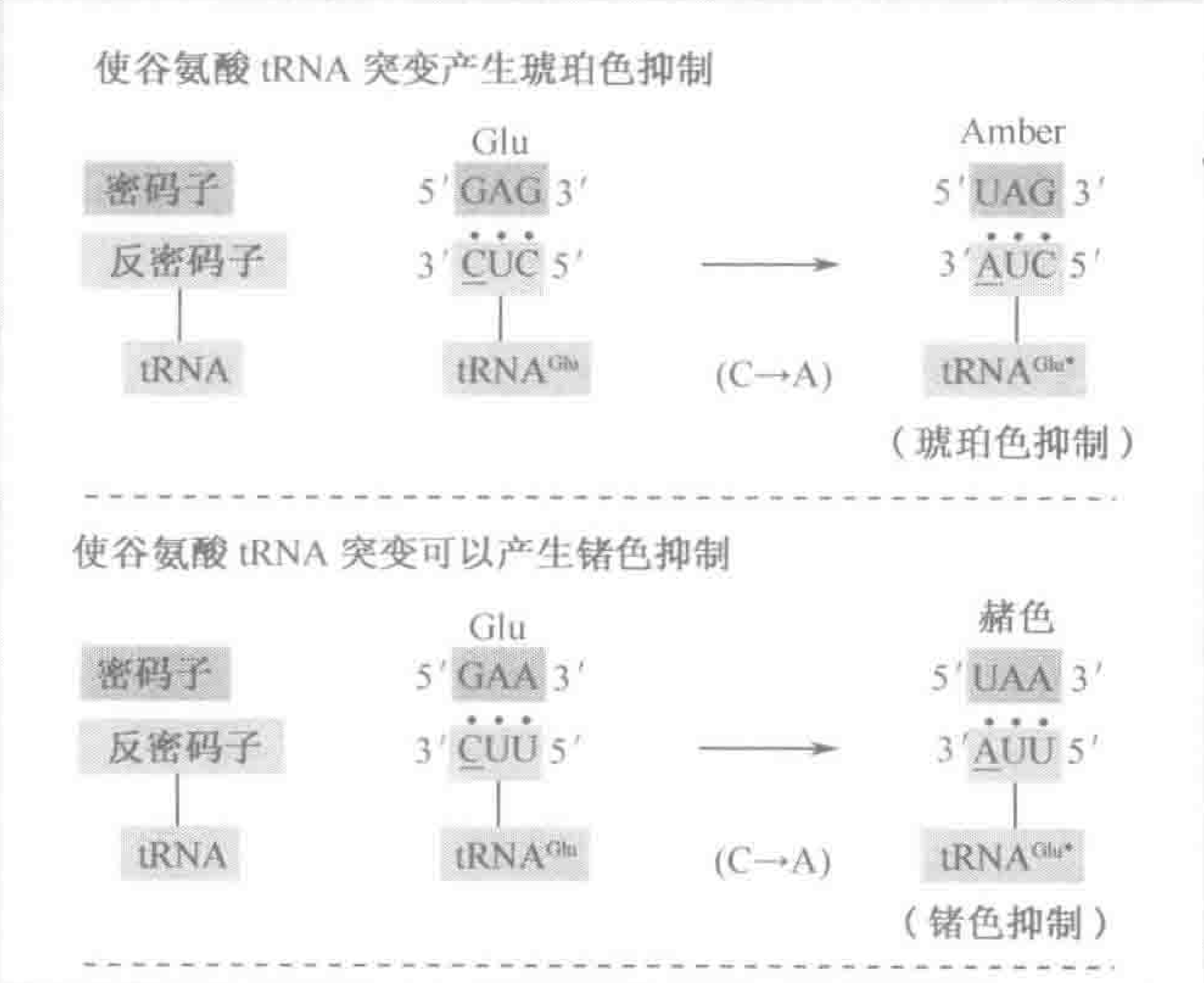
一般性的重组体 DNA 筛查通常取决于**插入失活**（insertional inactivation）：设计载体含有某种标记基因，赋予易于识别的某种表型以及含有独一无二的限制位点，使外源 DNA 能够插入标记基因，使其失活并改变表型。为达到这一目的，通常在标记基因内插入一**多克隆位点多接头**（ multiple cloning site polylinker ）修饰标记基因，多克隆位点多接头设计为含有特异限制性核酸酶的识别序列的双链寡核苷酸（这些酶预先存在的限制位点会从载体中被删除，如果必要的话，以确保单一克隆位点的存在）。

因为多接头短（约 30bp），且长度是三个核苷酸的倍数（维持标记基因的读框），所以它不影响标记基因的表达。然而，当外源 DNA 片段被克隆导入多接头时，标记基因可以有大片段插入，引起标记基因失活。常用的体系包括：

▶ **β 半乳糖苷酶筛查**　就标记物 β 半乳糖苷酶基因而言，在 Xgal 存在的情况下，插入失活导致细胞无色，而含有非重组体的细胞则变成蓝色：

框 5.3　无义抑制突变

tRNA 反密码子的碱基改变可以使其能够插入一对终止密码子反应的氨基酸。谷氨酸有被两种不同的 tRNA 分子识别的两个密码子，GAG 和 GAA。$tRNA^{Glu}$ 顶部携带识别谷氨酸密码子 GAG 的 CUC 反密码子。tRNA 基因突变可以产生突变的 $tRNA^{Glu}$，其反密码子的 3′碱基有 C→A 改变。这个突变的 tRNA 现在可以识别琥珀色终止密码子 UAG，并通过插入谷氨酸抑制琥珀色终止信号。下面的例子阐明了类似的适用于其他 $tRNA^{Glu}$ 反密码子的 3′碱基的（C→A）突变。这次产生的突变 tRNA 可以抑制赭色终止密码子的作用。



▶ **抑制型 tRNA 的筛查**　**抑制型 tRNA**（suppressor tRNA）基因是突变的 tRNA 基因，其携带突变的反密码子序列与正常终止密码子之一互补：UAA（赭色），UAG（琥

珀色）或 UGA（蛋白色）。为了与相关的终止密码子反应，抑制型 tRNA 插入一氨基酸（框 5.3）。宿主细胞典型地携带终止密码提前的有缺陷的标记基因，产生易于识别的表型。如果载体携带合适的抑制型 tRNA 基因来抑制标记基因突变，野生型表型将得到恢复。外源 DNA 克隆导入抑制型 tRNA 基因引起插入失活，恢复突变表型。

通过利用杂交和 PCR 直接筛查重组体

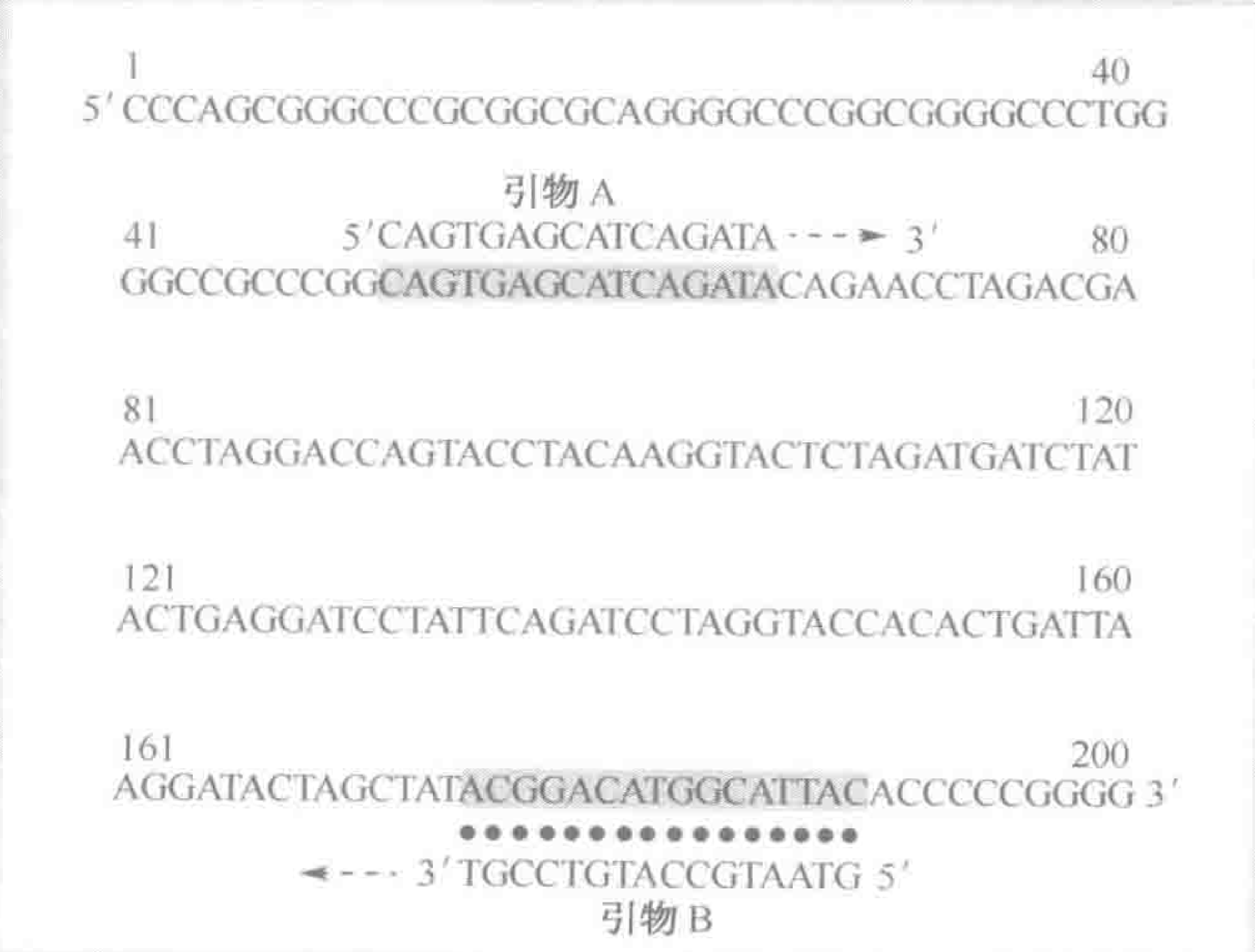
为了检测细胞中早先鉴定的 DNA 序列或与已知 DNA 序列相关的 DNA 序列，通常直接进行 DNA 文库筛查。例如，可以筛查携带非常大的克隆插入的 DNA 文库以得到非常大的重组体克隆用于功能分析，或筛查来自研究很少的种属的 DNA 文库以得到与已知人类基因相关的序列。这可以借助 DNA 杂交筛查或 PCR 完成：

- **杂交筛查** 用某种方法标记感兴趣的特殊 DNA 克隆，然后，作为杂交探针识别含有目的序列的细胞集落（节 6.4.1）；

框 5.4 序列标签位点（STS）的重要性

序列标签位点是简单重要的绘图工具，因为 PCR 能非常方便地检测序列是否存在。大多数 STS 是非多态的，在已经测序的基因组中，精确地知道每个 STS 的唯一亚染色体位点。STS 如何从 DNA 序列发展起来的例子如下所示。

在复杂基因组中，如人类基因组，16 核苷酸长的引物偶然地与相关而不相同的序列，但不是目的靶序列结合的机会并非无关紧要。然而，两个引物与非目的相关序列结合的机会正常情况下是非常低的。这两个相关序列碰巧是紧密相邻且在合适的方向上。在琼脂糖凝胶上分离扩增产物大小就可简单地检测反应的特异性。如果存在预期近似大小的单一大量 PCR 产物（上面例子中的 141 bp），那么就有非常好的机会，即对目的靶序列的检测是特异的，因此定义为 STS。



- **PCR 筛查** 只要 DNA 序列是已知的，通常可设计特异 PCR 试验来检测其是否

存在。人们已经认为该序列被标记（因为它可能总是被特异 PCR 试验识别）。如果序列发生在目的基因组内的唯一位点（位置），则此位点称为**序列标签位点**（sequence tagged site，STS）（框 5.4）。序列标签位点在制作基因组物理草图方面非常有用（节 8.3.2），但同样的原理可以用于文库筛查方面。在这种情况下，上千个单细胞克隆被储存，从每个细胞中分离的 DNA 重组沉积到多个微量滴定平皿孔内。来自不同孔的克隆池——按不同等级排列——可以检测特异的 STS 是否存在，以鉴定一个孔中含有目的 DNA 以及此后的原始细胞克隆（图 5.11）。

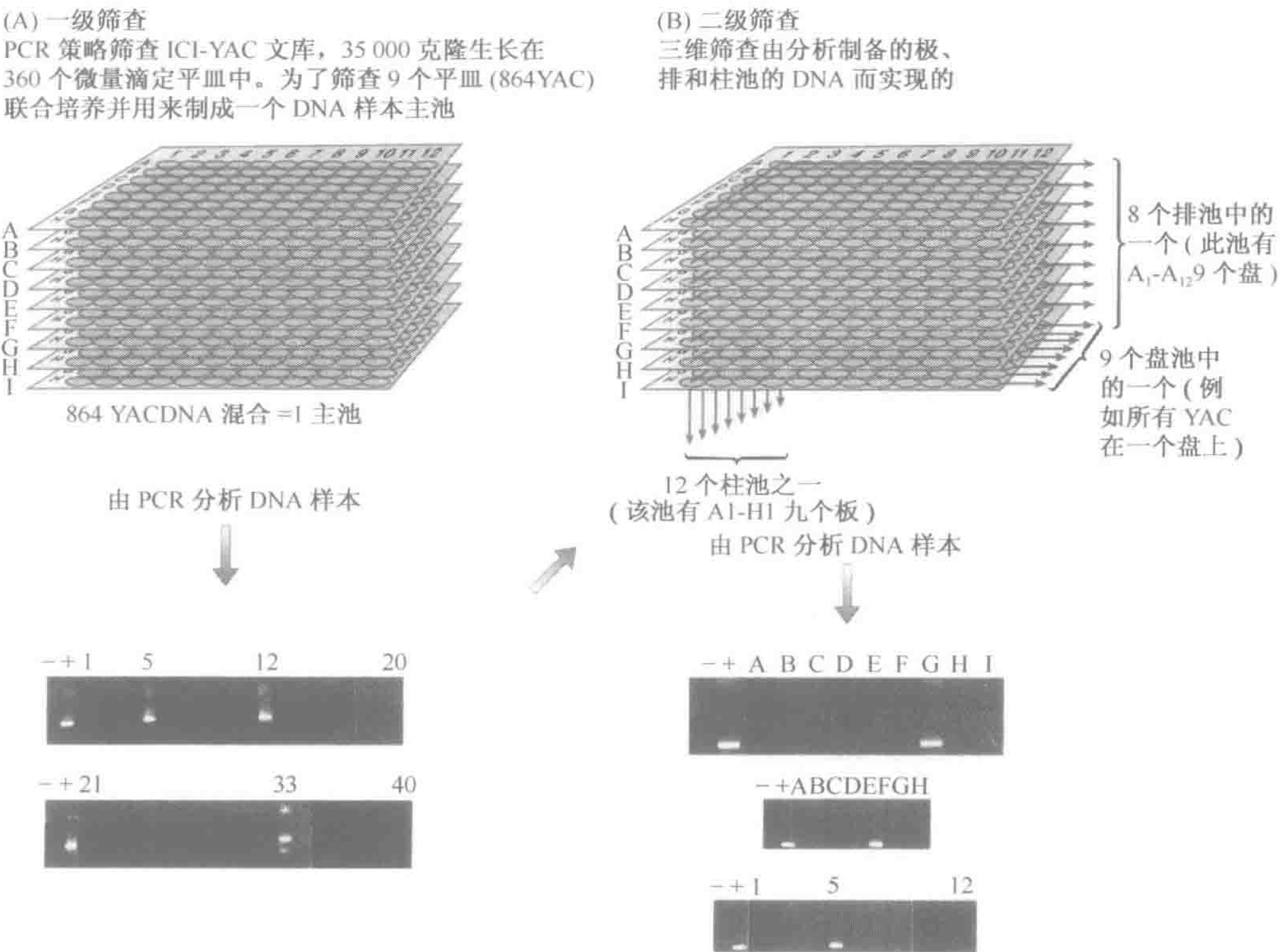


图 5.11　基于 PCR 的文库筛查

此例说明人类 YAC 文库的筛查。大约 35 000 个单克隆单独加入 360 个微量滴定平皿的 96 个孔中。为了便于筛查，将 9 个微量滴定平皿（板 A-I）中的所有 864 个克隆结合产生总共 40 个**主控池**（master pool）。由 Jones 等（1994）修改，Genomics，24，pp. 266～275，经 Academic Press 允许。(A) **一级筛查**（primary screening）涉及 40 个主控池的 PCR 检测。此例中，当参照阳性（+）和阴性（−）对照池 5，12，33 时，三个主控池呈阳性。(B) **二级筛查**（secondary screening）通过在阳性主控池中检测 864 个 YAC 的不同亚群识别单一 YAC，此例中为主控池 12。三维筛查 9 个**板池**（plate pool）（每个含 96 个 YAC），8 个**排池**（row pool）（每个含 106 个 YAC），12 个**柱池**（column pool）（每个含 72 个 YAC）中的每一个池，并在板 12G（顶层嵌板），排 E（中间嵌板），柱 5（底部嵌板）识别一个阳性 YAC。此处这个例子涉及含有无个性特征的 X 染色体序列的 YAC 的筛查。源自 Jones 等（1994），Genomics. 24（1），266～275，Elsevier 知情同意。图片由 University of Newcastle upon Tyne Sandie Herrell 博士惠赠。

5.4 扩增不同片段大小的克隆体系

表 5.2 利用不同克隆载体获得的常见插入 DNA 的大小

克隆载体	插入大小
标准高拷贝质粒载体	0～5kb
λ噬菌体插入载体	0～10kb
λ噬菌体替代载体	9～23kb
黏粒载体	30～44kb
噬菌体 P1	70～100kb
PAC（P1 人工染色体）载体	130～150kb
BAC（细菌人工染色体）载体	达到 300kb
YAC（酵母人工染色体）载体	0.2～2.0kb

细胞 DNA 克隆作为一种工具已经被广泛使用，主要制备大量纯 DNA 用于单个基因，基因簇或其他目的 DNA 序列的物理特性和功能研究。然而，不同的目的 DNA 序列大小可能差异很大（例如已知人类基因大小在 0.1kb～ 2Mb 之间不同）。第一个发展起来的细胞克隆体系只能克隆相当小的 DNA 片段。然而，最近克隆体系发展迅速，可以克隆非常大的 DNA 片段（表 5.2）。

5.4.1 标准质粒载体为在细菌（和简单真核）细胞中克隆 DNA 小片段提供一简单方法

为了使天然质粒分子适合作为克隆载体，通常进行几种修饰：

- 插入**抗生素耐受基因**（antibiotic resistance gene） （能够筛查载体是否存在——节 5.3.5）；
- 插入含有多克隆位点多接头的**标记基因**（marker gene）（使能够筛查重组体——节 5.3.5）。

例如，质粒载体 pUC19 含有一个具有用于多个限制性核酸酶的单克隆位点的多接头以及氨苄青霉素耐受基因，允许识别转化细胞（图 5.12）。此外，通过 β 半乳糖苷酶基因成分的插入失活筛选重组体，该基因的互补部分由特殊修饰的大肠杆菌宿主细胞提供。

5.4.2 λ 和黏粒载体为在细菌细胞中克隆中等大小 DNA 片段提供了有效的方法

质粒载体的主要缺点是接纳大片段 DNA 的能力严重受限：大多数插入片段是几个 kb 长，很少超过 5～10 kb。另外，用质粒载体转化细菌细胞的常规方法相对来说是无效的。为了解决这些难题，人们在早期的注意力主要集中在使用 λ 噬菌体作为克隆载体的可能性。野生型 λ 病毒颗粒含有一约 50 kb 的线状双链 DNA 的基因组，包装在蛋白外壳内，已经演变成高效感染 *E. coli* 细胞的机制。

λ 病毒颗粒附着于细菌细胞后，丢弃外壳蛋白，λDNA 被注入细胞。λDNA 末端是 12 个核苷酸长并与碱基序列互补的 5′突出末端。由于这些大的 5′突出末端可与碱基配对，所以它们是有效的黏性末端，与某些限制性核酸酶产生的小的黏性末端类似，但黏性更高。如此的黏附特性赋予该序列名为——**黏序列**（cos sequence ）。一旦进入细菌细胞，黏序列进行碱基配对，通过细胞连接酶封闭缺口，导致环状双链 DNA 形成。之后，λDNA 可以进入两条可选择的途径（图 5.13）：

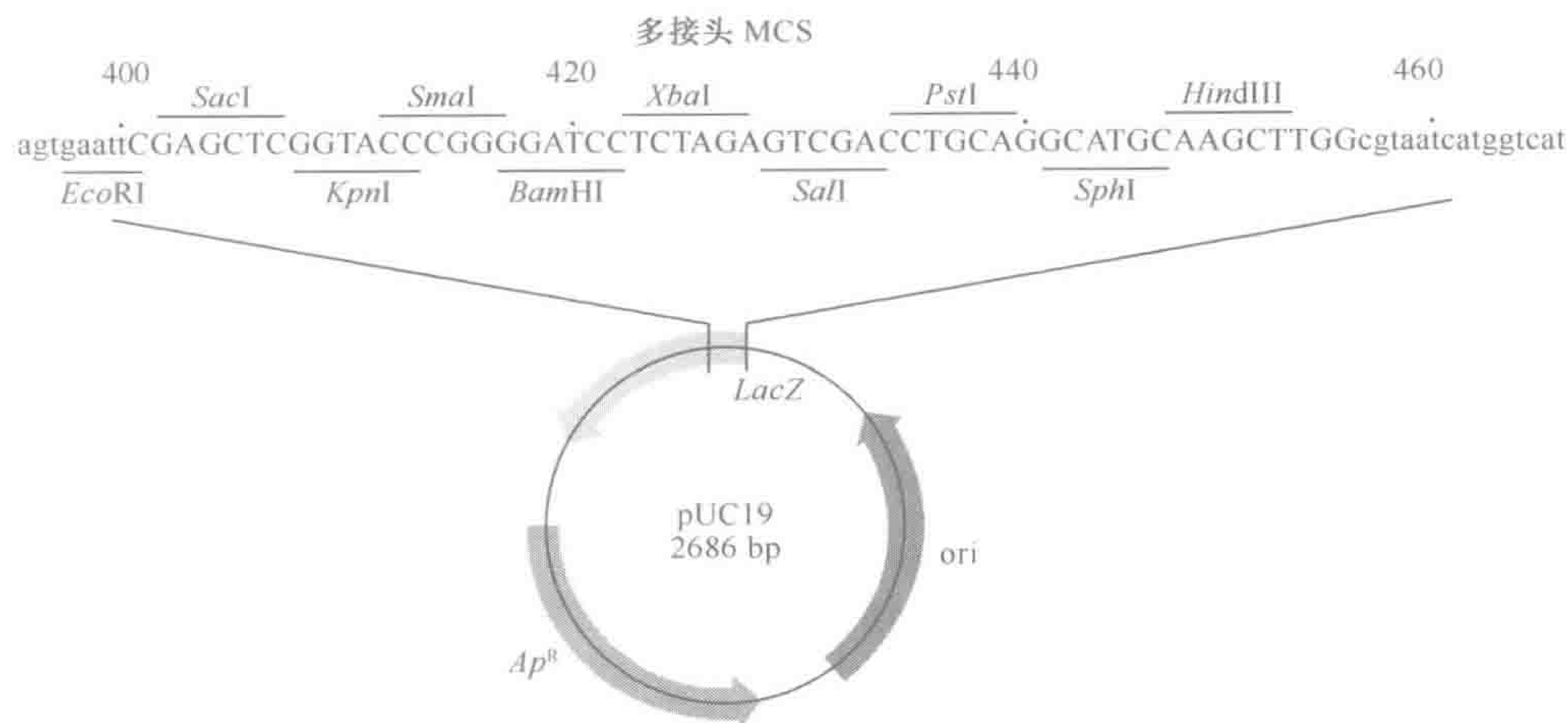


图 5.12　质粒载体 pUC19 的图谱

复制起点（ori）最初起源于 ColE1 样质粒，pMB1。氨苄青霉素耐受基因（Apr^R）允许筛选含有载体分子的细胞。部分 *lacZ* 基因包括在内，表达 β 半乳糖苷酶的氨基末端片段。这与宿主细胞中突变的 *lacZ* 基因互补：尽管单个存在无活性，但载体和宿主细胞 *lacZ* 序列产物可以结合形成功能产物。54bp 多接头多克隆位点（大写字母）以这种方式插入载体 *lacZ*（小写字母）成分中，以保留可读框和功能表达。然而，插入多克隆位点（MCS）的克隆会引起插入失活和 β 半乳糖苷酶活性的丧失。

- **裂解周期**（lytic cycle）　DNA 起初是双向的，随后以滚动循环模式进行复制，产生单位长度的线状多聚体。外壳蛋白合成且在 cos 位点切除 λ 多聚体，产生包装在蛋白外壳内的单位长度的基因组。某些 λ 基因产物裂解宿主细胞，使病毒逃逸并感染新的细胞；
- **溶源化状态**（lysogenic state）　λ 基因组具有基因 *att*，该基因在 *E. coli* 染色体中有同源序列。两个 *att* 基因的并列可能导致 λ 和 *E. coli* 基因组之间重组，以及随后在 *E. coli* 染色体内的 λDNA 的整合。在这种状态下，λDNA 被称为**前病毒**（provirus），宿主细胞叫**溶源菌**（lysogen），因为尽管 λDNA 可以长期稳定地整合，但是它有从宿主染色体切掉并进入裂解周期的能力（图 5.13）。需要溶源化功能的基因定位在 λ 基因组的中心部分（图 5.14）。

是否进入裂解周期还是溶源化状态的决定是由两个调节基因调控的，*c*I 和 *cro*。这两个基因相互拮抗：在裂解状态下，*cro* 蛋白占优势，导致 *c*I 的抑制。而在溶源化状态下，*c*I 抑制因子占优势，抑制包括 *cro* 在内的其他 λ 基因的转录。在正常生长的宿主细胞中，有利于溶源化状态，λ 基因组伴随宿主染色体 DNA 一起复制。对宿主细胞的损害有利于宿主细胞过渡到裂解周期，使病毒能够逃逸受损害的细胞，感染新的细胞。

为了设计合适的 λ 克隆载体，有必要设计外源 DNA 能够被连接到体外的 λ 复制子且合成的 DNA 重组能够高效转化 *E. coli* 细胞体系。通过发展体外包装系统完成高效转化。外包装系统模仿野生型 DNA 包装在蛋白外壳内的方式，导致高效转染（图 5.15）。

下面描述了通过修饰 λ 噬菌体或利用 cos 序列施加的筛选片段大小发展起来的几个主要类型的克隆载体。

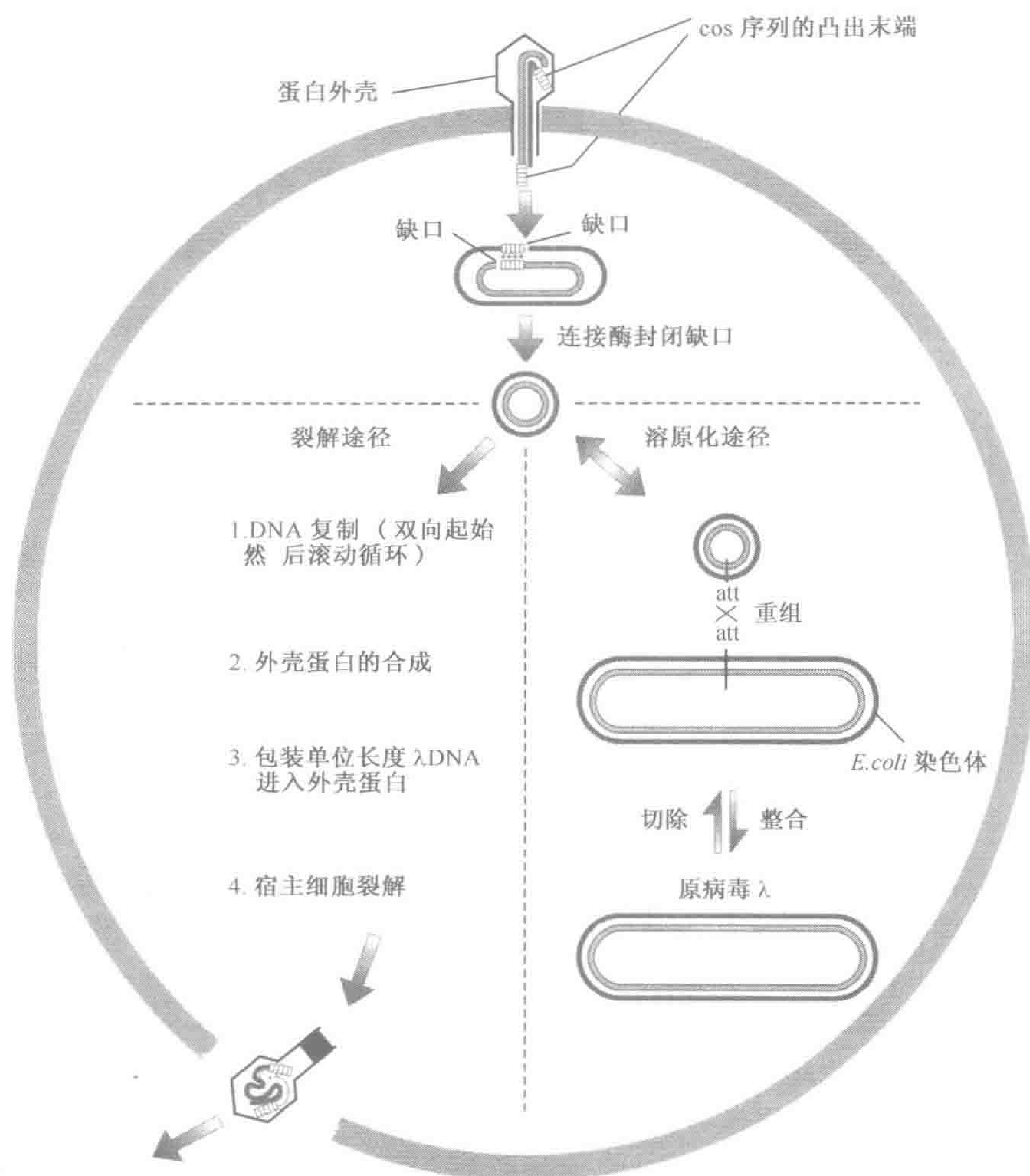


图 5.13　λ 噬菌体可以进入裂解和溶源化两种途径

- **置换型 λ 载体**（Replacement λ vector）　只有 37～52kb 长的 DNA 分子可以稳定地包装进入 λ 颗粒。λ 基因组的中心部分含有溶源化周期需要的但不是裂解功能必需的基因。结果，此部分可以切除并由外源 DNA 片段替代。应用此策略，有可能克隆长达 23kb 的外源 DNA，这样的载体通常用于制备基因组 DNA 文库；
- **插入型 λ 载体**（insertion λ vector）　用于制备 cDNA 文库的 λ 载体不需要有插入大片段的能力（大多数 cDNA <5kb）。插入载体的设计经常涉及 λ 基因组的修饰，允许插入克隆进入 cI 基因；
- **黏粒载体**（cosmid vector）含有插入小质粒载体中的 cos 序列。大的（约 30～445kb）外源 DNA 片段可以在体外包装反应中利用这样的载体被克隆，因为总的黏粒载体大小通常大约是 8 kb（图 5.16）。

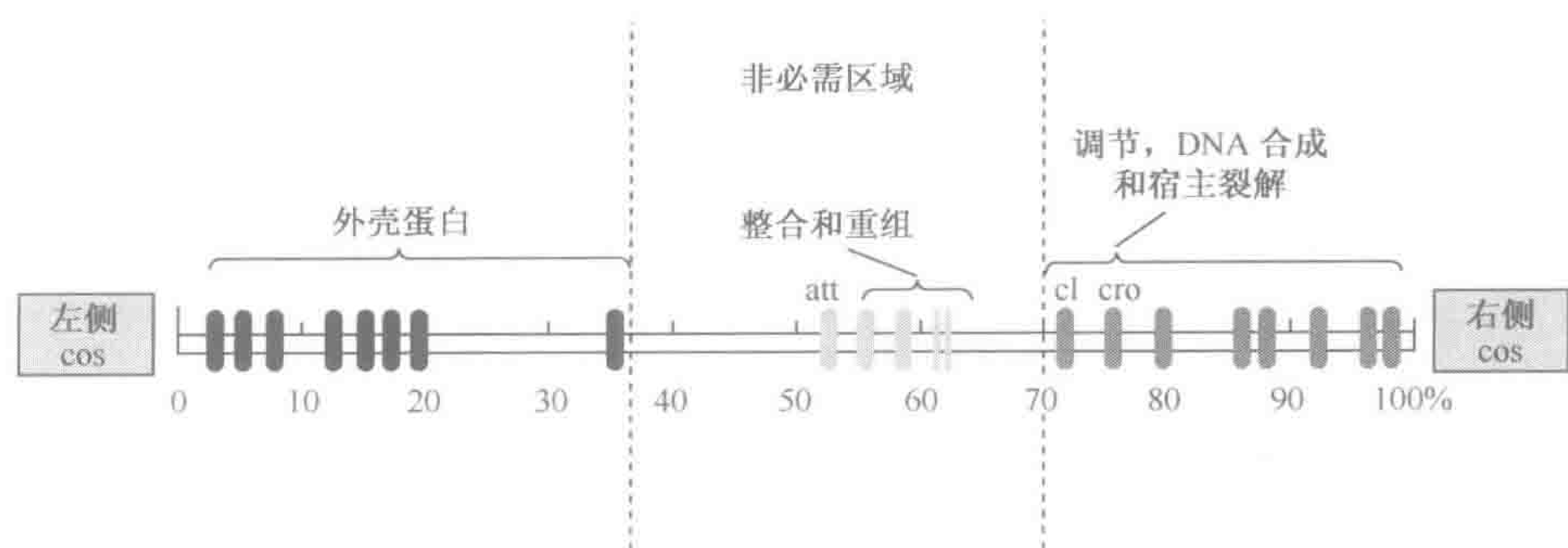


图5.14　λ基因组图谱，显示基因位置（竖条）

在λ置换型载体中，限制性内切核酸酶消化切除非必需区域，保留λ左臂和λ右臂。外源DNA片段可以连接到两臂，替换原来的"填塞"片段，提供最大的超过20 kb大小的插入片段。

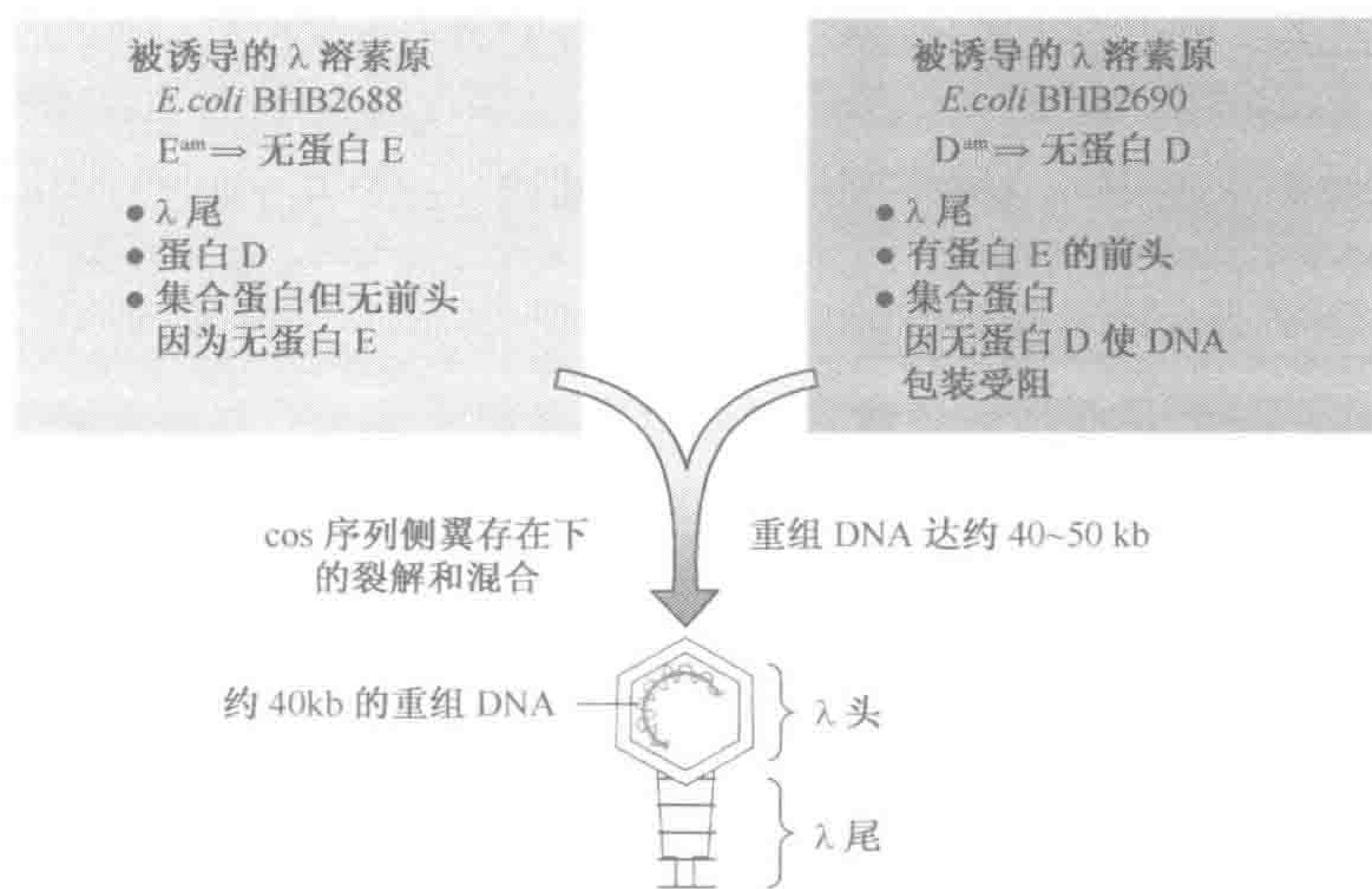


图5.15　利用两个突变λ溶源菌的混合裂解液可以在λ噬菌体外壳蛋白内进行体外DNA包装

λDNA的正常体内包装首先与制备由基因*E*编码的主要外壳蛋白组成的结构**前头**（pre-head）有关。一个λDNA单位长度插入前头，单位长度是由邻近cos位点断裂制备的。然后，次要外壳蛋白D插入前头从而完成头的成熟，同时其他基因产物充当**集合蛋白**（assembly protein），确保完整的尾与完整的头连接。由导入基因*E*（E^{am}）的琥珀色突变引起制备的蛋白E缺陷，阻止BHB2688引起的前头形成。基因*D*（D^{am}）的琥珀色突变阻止了携带密闭DNA的前头成熟为完整的头。然而，BHB2688/BHB2690混合裂解液成分补充了相互的缺陷，为所有产物提供了正确的包装。

5.4.3　大的DNA片段可以利用噬菌体P1和F因子质粒载体在细菌细胞中克隆

细菌人工染色体（BAC）载体

细菌细胞中许多用于DNA克隆的载体取决于达到中等拷贝数目的复制子。高拷贝数目产生高产量的DNA：载体分子繁殖的每个细胞会有几个拷贝到许多拷贝的载体分子。一个致命的缺点是这样的载体经常显示插入结构的不稳定性，导致克隆的DNA部

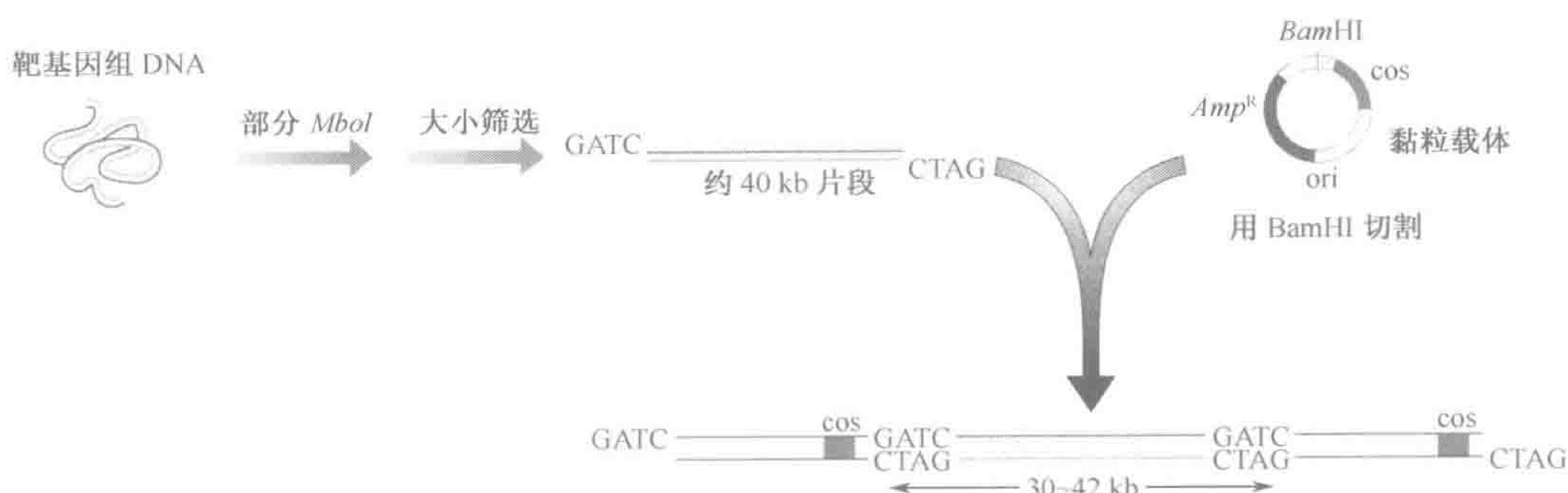


图 5.16 与被切割的黏粒载体分子连接可以制备载体-靶多连体，通过 cos 序列产生旁侧大的外源 DNA 片段

分的缺失或重排。这样的不稳定性在重复序列频发的真核细胞起源的 DNA 插入的情况下尤为常见。结果，难以在细菌细胞中克隆和维持完整大的 DNA。

为了克服这一缺陷，近来，人们的注意力主要集中在低拷贝数目的复制子载体，如 *E. coli* 发育质粒和 F 因子。这种质粒含有两个基因 *parA* 和 *parB*，它们可以在每个 *E. coli*细胞维持 F 因子的拷贝数目 1～2 个。F 因子体系的载体能够接纳大的外源 DNA 片段（>300 kb）。所产生的重组体可以利用**电穿孔**（ electroporation ）（一种将细胞暴露于高伏特电压，目的是使细胞质膜选择性渗透增加的方法）高效率转导进入细菌细胞。然而，因为合成的**细菌人工染色体**（BAC）含有低拷贝数目复制子，所以只有低产量的重组体 DNA 可以从宿主细胞中得到恢复。

噬菌体 P1 载体和 P1 人工染色体（PAC）

某些噬菌体有相对大的基因组，因此，有潜能发展成可以容纳大的外源 DNA 片段的载体。**噬菌体 P1**（bacteriophage P1）就是这样的载体，像 λ 噬菌体一样，在蛋白外壳内包装自身的基因组；110～115 kb 的线状 DNA 被包装在 P1 蛋白外壳内。因此，已经设计 P1 克隆载体的 P1 成分包含在环状质粒中。

P1 质粒载体可以被切割，产生两个载体臂，达 100kb 的外源 DNA 在体外可以连接到臂上并包装进入 P1 蛋白外壳。重组的 P1 噬菌体允许吸附到合适的宿主，随后重组的 P1 DNA 注入细胞，环化和扩增（Sternberg，1992）。人通过基本的 P1 克隆系统，改进接受的插入片段大小范围一直是应用携有 P1 载体的 T4 噬菌体体外包装系统，该 T4 系统能够恢复插入的大小达 122kb。最近，P1 和 F 因子体系的特性已经结合起来制备克隆体系（Iouannou *et al*.,1994）。

5.4.4 酵母人工染色体（YAC）能克隆 Mb 片段

最广泛使用的克隆巨大 DNA 片段的体系与**酵母人工染色体**（YAC；Schlessinger，1990）的构建有关。某些真核细胞的序列，特别是那些具有重复序列结构的序列，是难以或不可能在没有这种类型 DNA 结构的细菌细胞中增殖的，但有希望在属于真核细胞的酵母细胞中增殖。然而，YAC 提供的主要优点是具有克隆非常大的 DNA 片段的能

力。如此的发展基于这样一种认识，即正常染色体功能不需要染色体中大量的 DNA。正如图 2.5 详述的那样，酵母染色体的基本功能成分分成三部分：

▶ **中心粒**（centromere）　有丝分裂期姐妹染色单体分离和第一次成熟分裂期同源染色体的分离需要中心粒；

▶ **端粒**（telomere）　为线性分子的完全复制和保护染色体末端免除核酶攻击所必须；

▶ **自主复制序列**（autonomous replicating sequence）　元件为染色体 DNA 的自主复制所必需并被认为作为特殊复制起点发挥作用。

在每一种情况下，酵母细胞内的功能活性所必须的 DNA 片段被限制到至多几百个碱基对（图 2.5）。结果，有可能把基于染色体复制子（自主复制序列元件）应用的新克隆体系看作染色体外复制子（在质粒和噬菌体中发现的那些复制子）的可选择方法，与人工染色体的构建有关。

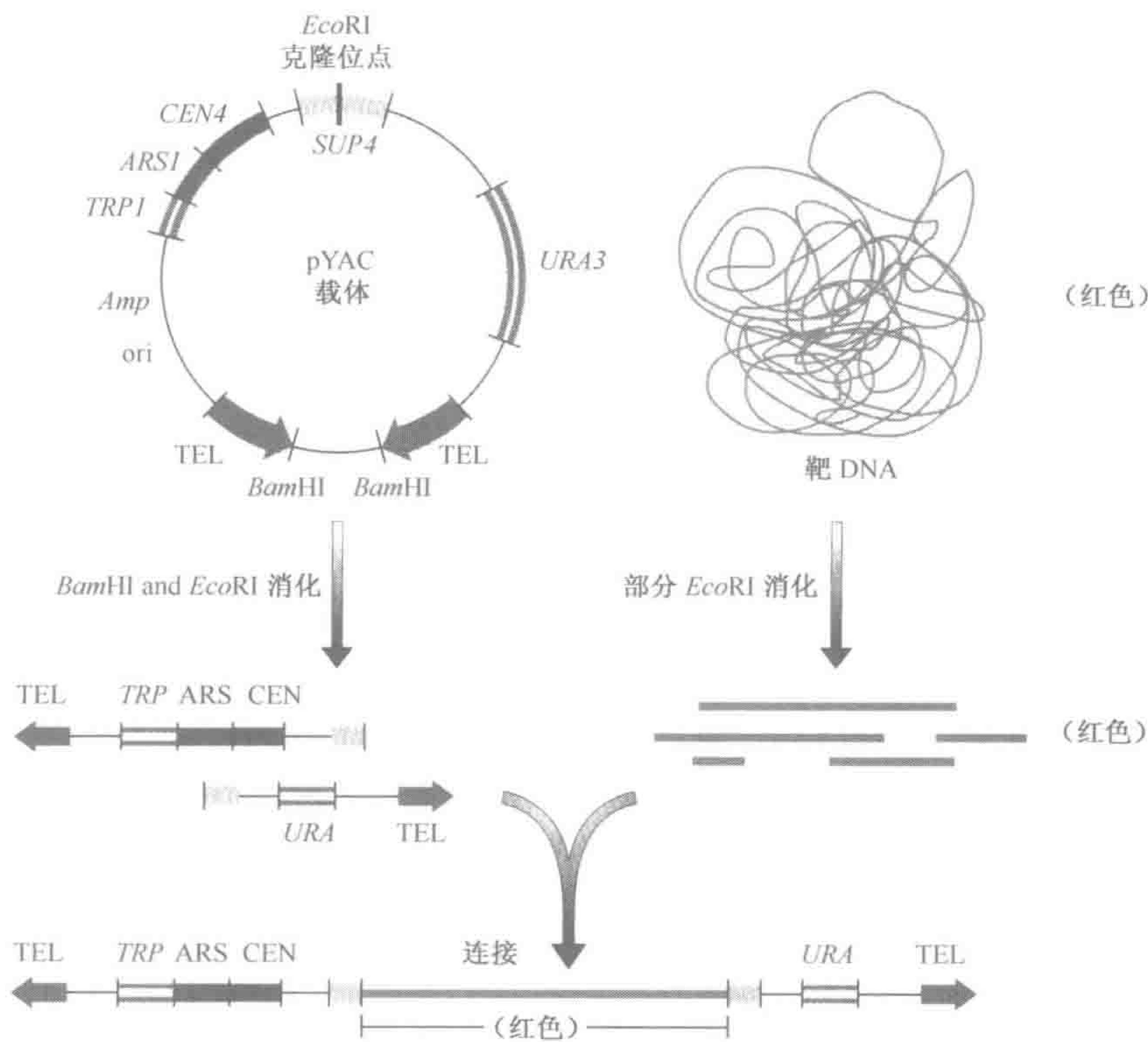


图 5.17　制备 YAC

载体 DNA 序列包括：CEN4，中心粒序列；TEL，端粒序列；ARS1，自主复制序列；*Amp*，氨苄青霉素耐受基因；ori，在 *E. coli* 宿主中繁殖的复制起点。载体与特殊酵母宿主细胞 AB1380 一起使用，后者是红色的，因为它在基因 *ade-2* 中携带赭色突变，与腺苷酸代谢有关，导致红色色素的蓄积。然而，载体携带 *SUP4* 基因，一种抑制型 tRNA 基因（框 5.3），该基因克服了 *ade-2* 的赭色突变，使其野生型活性恢复，产生无色集落。宿主细胞也设计含有隐性 *trp1* 和 *ura3* 等位基因，它们与载体中对应的 *TRP1* 和 *URA3* 等位基因互补，为鉴别含有 YAC 载体的细胞提供了筛选体系。克隆外源 DNA 片段进入 *SUP4* 基因引起抑制基因功能的插入失活，恢复突变（红色）表型。

欲制备 YAC，需要在酵母细胞中简单地重组可以行使功能的四个短序列：两个端粒，一个中心粒和一个 ARS 元件，与合适的外源 DNA 片段结合起来产生线性 DNA 分

子，其中端粒序列恰好位于末端（图 5.17）。所形成的重组体不能直接转染进入酵母细胞。取而代之，酵母细胞不得不以这种方式处理，以便除掉细胞外壁。形成的酵母**原生质球**（spheroplast）可以接纳外源片段，但渗透性不稳定并需要包埋在琼脂中。整体的转化效率非常低并且克隆的 DNA 产量低（大约每个细胞一个拷贝）。然而，克隆大的外源 DNA 片段的能力已经使 YAC 成为物理作图的重要工具（节 8.3.2）。

5.5 制备单链、诱变 DNA 的克隆体系

单链 DNA 克隆对于包括 **DNA 测序**（DNA sequencing）（因为获得的序列更清楚和更易读）和**位点专一诱变**（site-directed mutagenesis）（在克隆 DNA 中的特异位点需要以精确、事先确定的方式被改变）在内的一些应用是有帮助的。如果功能检测对于目的 DNA 是可用的，可以设计位点专一诱变产生特异寡核苷酸替代、缺失等等，可以帮助识别关键氨基酸残基或其他功能重要的序列。

5.5.1 用于 DNA 测序的单链 DNA 可以利用 M13 或噬粒载体或线性 PCR 扩增而获得

单链 DNA 重组克隆通常作为模板利用载体而用于 DNA 测序，载体是在其生活周期的某个阶段自然接纳单链 DNA 形式的某种噬菌体载体。因为载体序列是已知的，所以很方便使用与邻近克隆位点的载体中序列互补的单一载体特异序列引物。

M13 载体

M13 是可以感染某些 *E. coli* 菌株的**丝状噬菌体**（filamentous bacteriophage）。其 6.4kb 的环状单链基因组被包装在蛋白外壳内，形成长丝状结构。吸附到菌株后，M13 基因组进入细菌细胞并转化成双链形式，即**可复制形式**（replicative form，RF），充当制备无数基因组拷贝的模板。在一定时期后，噬菌体编码产物启动 DNA 合成朝向单链生成的方向，然后移动到细胞膜。在这里，它们被包装在蛋白外壳内，上百个成熟噬菌体从感染细胞中逸出，而没有细胞裂解。M13 载体是在携带多克隆位点的可复制形式的基础上建立起来的，可以接纳限制大小的外源插入。后者可以导入合适的 *E. coli* 菌株。在一定时间后，收获噬菌体颗粒并剥去蛋白外壳以释放单链 DNA 重组，直接作为 DNA 测序反应中的模板（图 5.18A）。

噬粒载体

丝状噬菌体基因组的小片段，如 M13（或相关的丝状噬菌体 fd 或 f1）可以插入**质粒**（plasmid）形成叫做**噬粒载体**（phagemid vector）的杂种载体。选择的噬菌体序列包含所有的 DNA 复制需要的顺式作用元件并组装进入噬菌体颗粒。它们允许成功地克隆几千个碱基长的插入片段（不像 M13，这样的插入是不稳定的）。含有重组噬粒的合适 *E. coli* 菌株转化后，细菌细胞被丝状**辅助噬菌体**（helper phage）超感染，如提供外壳蛋白的 f1。超感染的细胞分泌的噬菌体颗粒是辅助噬菌体和重组噬粒的混合物（图 5.18B）。混合的单链 DNA 群体可以直接用于 DNA 测序，因为始动 DNA 链合成的引物设计与邻近克隆位点的噬粒序列特异地结合。普遍使用的噬粒载体包括 pEMBL 质粒系列和 pBluescript 家族（图 5.18B）。

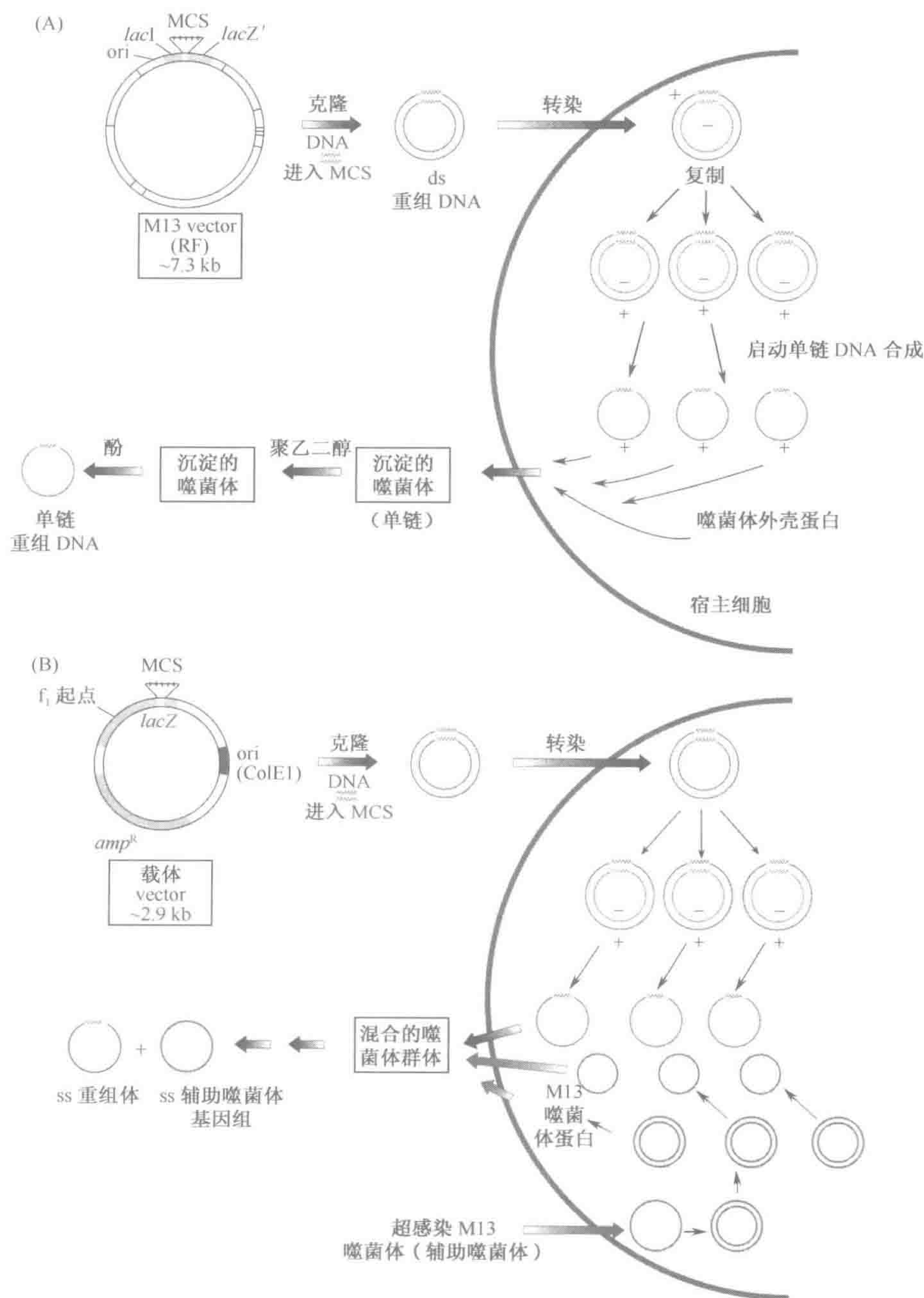


图 5.18　应用 M13 和噬粒载体制备单链 DNA 重组

（A）**M13 载体**　M13 载体是含有 *lacZ* β 半乳糖苷酶体系非功能成分的 M13 衍生物的可复制形式（RF），*lacZ* β 半乳糖苷酶体系通过与 *E.coli* JM 系列中存在互补的 *lacZ* 成分在功能上可以互补。启动单链 DNA（仅＋链）产生之前，双链 M13 DNA 重组进入 DNA 复制的正常周期，产生许多基因组拷贝。成熟的重组噬菌体离开细胞而没有裂解它。（B）**噬粒载体**（phagemid vector）pBluescript 系列质粒载体含有两个复制起点：正常的一个来自 *Co*/E1，第二个来自 f1，在丝状噬菌体基因组存在的情况下，*Co*/E1 和 f1 将特化生成单链 DNA。携带 M13 噬菌体的转化细胞的超感染导致两种类型的噬菌体样颗粒从细胞中释放：原有超感染噬菌体和噬菌体蛋白外壳内的质粒重组体。对噬粒载体特异的测序引物被用于获得明确的序列。

线性 PCR 扩增

一种称为**循环测序**（cycle sequencing）的 PCR 测序形式（框 7.1）利用修饰的 PCR 反应产生测序的单链模板。只用一个引物就可以完成线性 PCR 扩增，这样单链产物以线性而不是在标准 PCR 中看到的指数扩增形式蓄积。

5.5.2　寡核苷酸错配诱变可能在克隆的任何基因中产生预期的单核苷酸改变

基因功能的许多体外检测目的在于获得编码多肽中有关单个氨基酸的重要信息。试图评价在已知致病基因中发现的特殊无义突变是致病的，抑或仅泛泛地评价一种特殊氨基酸对蛋白质生物学功能的作用是恰当的。

一种普遍流行的方法与克隆基因或 cDNA 进入 M13 或噬粒载体来获得单链 DNA 重组有关（见前节）。设计诱变的寡核苷酸引物，其序列与只是单个碱基差异的突变区域的基因序列完全互补：在预期的突变位点，它产生了与预期的突变核苷酸而不是原来核苷酸序列互补的碱基。然后使诱变的寡核苷酸引导 DNA 合成，产生包含预期突变的全长互补序列。新形成的异源双链用于转化细胞，且可以通过筛查突变鉴定预期的突变基因（图 5.19）。

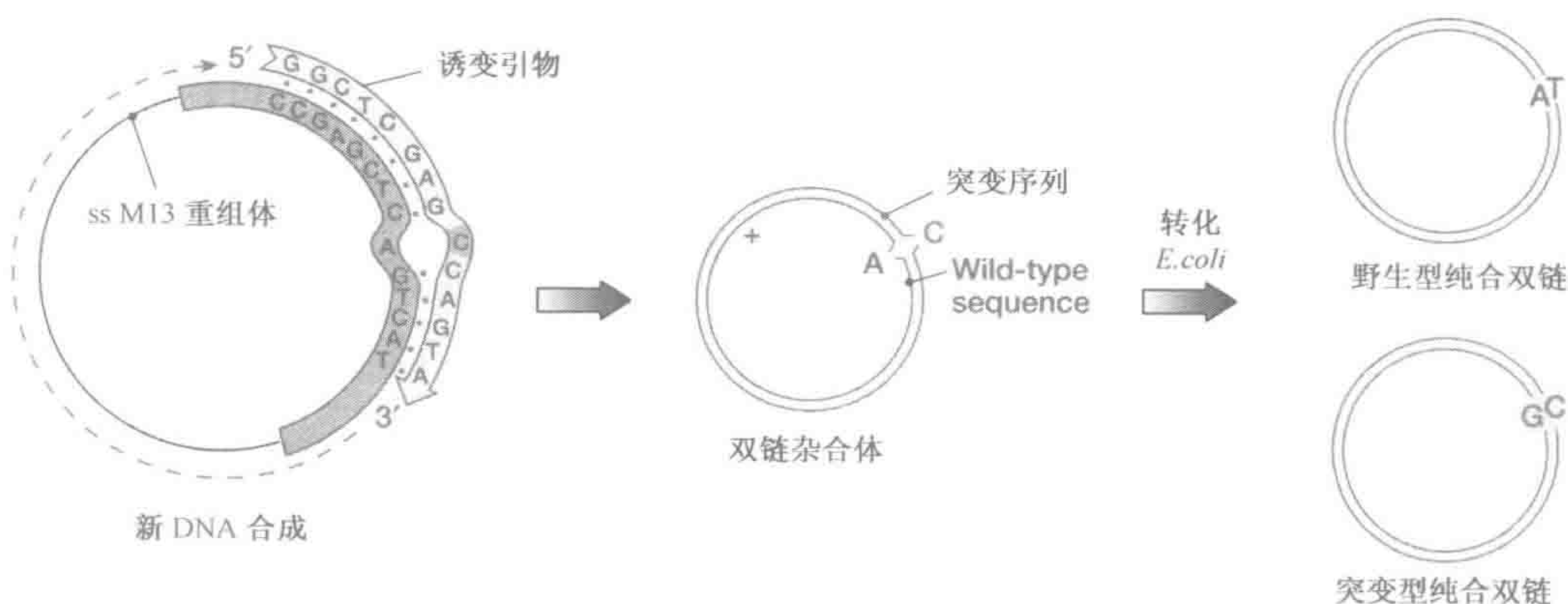


图 5.19　寡核苷酸错配诱变可以在克隆分子内预先确定的单一位点产生目的点突变

图仅仅阐述了许多细胞寡核苷酸错配诱变不同方法中的一种。例子说明利用诱变的寡核苷酸以指导在一基因中单核苷酸替代。基因被克隆进入 M13 以便产生单链 DNA 重组。设计寡核苷酸引物与该基因的部分序列互补，后者含有要突变的核苷酸（A）并在那个位点含有预期的非互补碱基（C，不是 T）。尽管内部错配，但诱变引物的复性是可能的，且第二条链的合成可以通过 DNA 聚合酶延伸，缺口由 DNA 连接酶连接。合成的异源双链可以转化进入 *E. coli* 细胞，于是可以获得重组体的两个群体：野生型和纯合双链。后者可以通过分子杂交（通过使用突变引物作为等位基因特异寡核苷酸探针；见图 6.11）或 PCR 等位基因特异扩增方法识别（图 5.4）。

除了单核苷酸替代外，其他小范围突变也可以被导入。例如，有可能导入三核苷酸缺失，引起编码的多肽除去一个氨基酸，或增加一个新的氨基酸的三核苷酸插入。假设诱变的寡核苷酸足够长，即使中心部分有大量的错配，它也能够与基因模板特异结合。利用**盒式诱变**（cassette mutagenesis）仍然可以导入较大的突变，在这种情况下，原基因的原序列的特殊区域被删除并由寡核苷酸盒式替代（Bedwell *et al*.,1989）。

5.5.3　PCR 诱变包括目的序列或化学基团与靶序列结合和位点专一诱变

PCR 的位点专一诱变已经变得愈来愈受欢迎并且已经设计各种策略能使碱基替代、缺失和插入成为可能（见下文以及 Newton and Graham，1997）。除了在靶 DNA 制备预先确定的特异突变，一种称为 5′附加诱变的诱变形式允许以同样的方式加入目的序列或化学基团，就像利用寡核苷酸接头连接的那样。

5′附加诱变（5′ add-on mutagenesis）是常用的一种方法，其中，一个新序列或化学基团加到引物设计的 PCR 产物的 5′端，引物的 3′部分有想要的特异序列而引物的 5′部分有携带附加的化学基团的新序列或一序列。附加的 5′序列不参加 PCR 反应的第一次复性步骤（只有引物的 3′部分对靶序列是特异的），但它随后被掺入扩增产物中，因此产生重组产物（图 5.20A）。各种受欢迎的附加 5′序列的选择方法包括：（Ⅰ）合适的限制位点，可以使后续的细胞克隆变得容易；（Ⅱ）一功能成分，例如驱动表达的启动子序列；（Ⅲ）含有报告基团或标记基团修饰的核苷酸，诸如生物素化的核苷酸或荧光素。

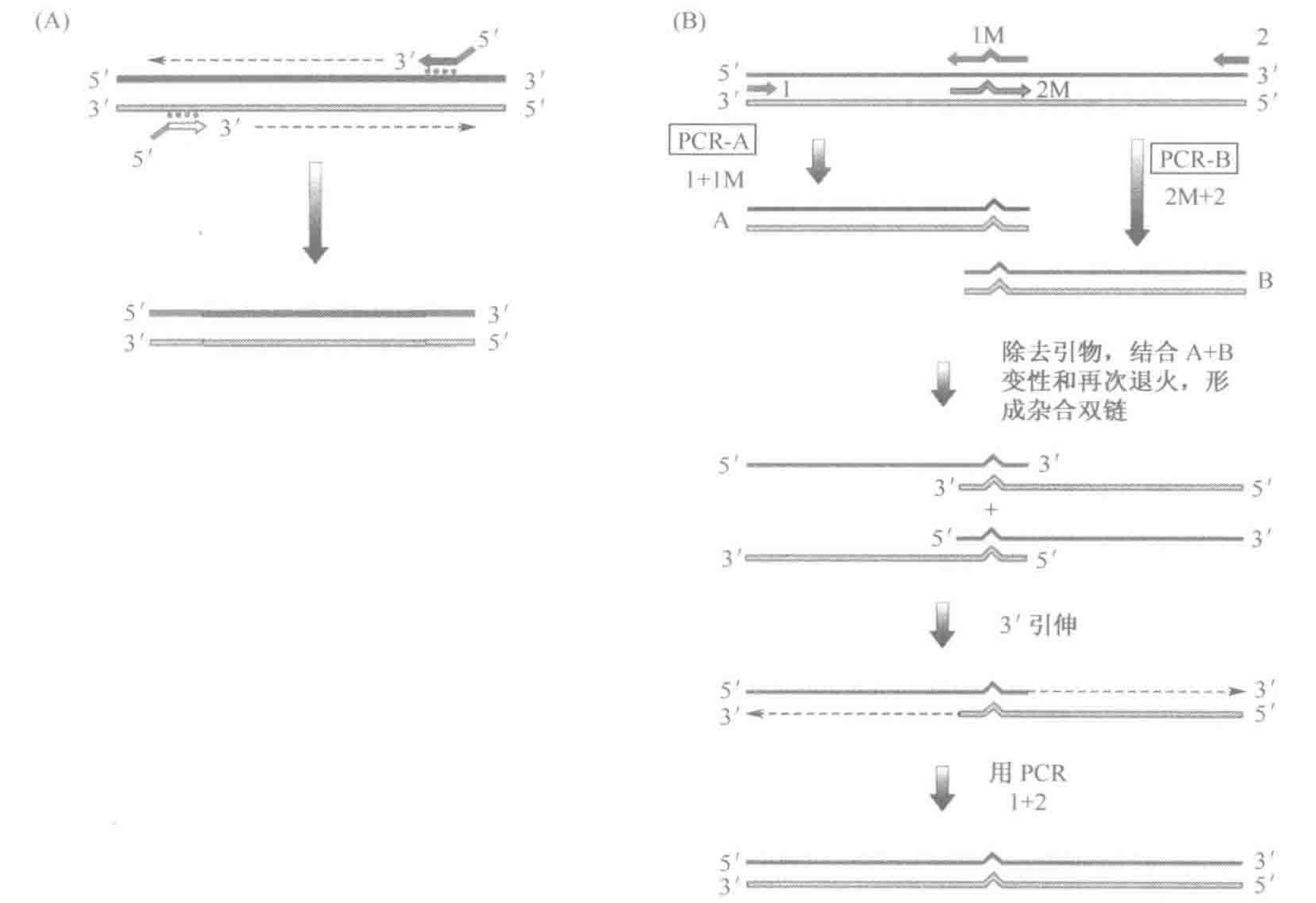


图 5.20　PCR 诱变

（A）**5′附加诱变**（5′add-on mutagenesis）引物可以在 5′端被修饰导入，例如，一标记基团（图 7.11）、一个序列含有合适的限制位点或噬菌体启动子驱动基因表达。（B）**位点专一诱变**（site-specific mutagenesis）显示的诱变可能产生携带位于中心片段的预先确定的特异突变的扩增产物。PCR 反应 A 和 B 被看作是扩增含有导入突变（通过使用突变引物：1M 或 2M 的精确碱基错配）的 DNA 的重叠片段。两个产物结合、变性和复性后，DNA 聚合酶可以延伸携带凹陷的 3′端的异源双链的 3′端。因此，通过只使用外部引物 1 和 2，携带中心片段的导入突变的全长产物可以被扩增。

错配引物的诱变（mismatched primer mutagenesis）设计引物仅与靶位点部分互补，但以这种方式引物仍会与靶序列特异结合。不可避免的这意味着突变被导入接近PCR产物的末端。正如图18.8所述，这种方法可以用来导入人工诊断的限制性位点，可以筛查已知突变。突变也可以应用错配引物在被选择序列的任何位点导入。设计两个诱变反应，其中，两个分开的PCR产物有含此突变的部分重叠序列。变性产物结合，在更中心的位置产生较大的携带突变的产物（Hguchi，1990；图5.20B）。

5.6 设计表达基因的克隆体系

当目的只是扩增导入的DNA以获得足够量用于各种后续结构和功能研究时，使用节5.4描述的克隆体系。然而，就基因序列而言，有许多情况不仅需要扩增和繁殖克隆的DNA，值得期待的是能够以某种方式表达该基因（**表达克隆**，expression cloning）。在每种情况下，需要由克隆体系提供合适的表达信号。应用PCR体系可以构建表达克隆，但通常利用细胞克隆体系进行。可以使用各种广泛的克隆体系，这取决于：

- **表达产物的类型** 就某些目的而言，能够获得某种RNA产物就足够了。例子包括用于组织原位杂交研究的反义RNA探针（核糖探针，图6.3）的制备，或制备反义RNA用于抑制或破坏特殊基因的表达，或功能研究或治疗目的（节20.2.6和节21.7.5）。然而，在许多情况下更期望蛋白产物的表达；
- **环境类型** 有时在体外表达该产物就足够了。然而，经常期待能够在非常明确的原核或真核细胞系的特殊细胞体系表达该产物；
- **表达体系的目的** 可以设计表达体系仅用于研究表达。然而，在许多情况下，目的是获得大量的表达产物，因为需要制备大量的特殊蛋白以帮助后续的晶体学研究或试图增加针对某种蛋白的特异抗体。

由于设计的特殊载体在特殊的宿主细胞类型中是有用的，所以许多不同表达克隆载体已经设计用于从细菌细胞至哺乳动物细胞范围内不同特异的宿主细胞系统。

5.6.1 大量的蛋白质可以通过在细菌细胞中的表达克隆制备

经常需要在真核细胞cDNA克隆表达载体中制备用于基础随访研究的医学相关化合物或蛋白质，例如结构研究。通常在这些情况下，cDNA仅设计提供特化的蛋白质序列的遗传信息和通过连接合适的强启动子、调节元件等等进入表达载体从外部提供表达信号。因为表达体系取决于DNA重组，并且也经常涉及人工融合蛋白或通过添加某些肽标记修饰的蛋白，所合成的这些蛋白质有时被称为**重组蛋白**（recombinant protein）。

细菌细胞的优点是生长迅速并且可以在培养基中容易扩增至非常大量的容积，而研究十分清楚的*E. coli*一直是异源（外源）蛋白表达最中意的宿主细胞（Baneyx，1999）。各种广泛的蛋白质可以利用噬菌体**T7 RNA聚合酶**（T7 RNA polymerase）的表达体系表达，能够产生来自几乎任何编码序列的完全转录物。

因为大量异源蛋白质的产生对于宿主细胞生长是有害的，甚至是有毒的，所以**可**

诱导启动子（inducible promoter）引起的表达控制是有益的。因为 *E. coli* RNA 不识别 T7 启动子，因此载体中被克隆的 DNA 在 T7 RNA 聚合酶缺乏的情况下很大程度上不表达。正常宿主 RNA 聚合酶不识别 T7 启动子，但出于克隆目的，使用细菌染色体内含有可诱导 T7 RNA 聚合酶的菌株。例如，当使用 pET 载体克隆体系时（图 5.21A），宿主通过 lac 启动子调节含有一 T7 RNA 聚合酶而被修饰，所以当要使用 *lac* 诱导剂异丙基-β-D-硫代半乳糖苷（IPTG）时，宿主可以被诱导。结果，转化的细胞可以被筛选，并且大量生长而无表达；之后，可以加入 IPTG 来诱导表达，而且不久即可收获细胞。

尽管细菌对于外源蛋白质表达有许多优点，但存在各种类型的限制：

- **翻译后加工**　对在细菌细胞中产生的某些真核细胞的蛋白质而言，缺乏正常的糖基化或磷酸化模式意味着蛋白质变得不稳定，或显示受限，抑或没有生物学活性；
- **蛋白质长度**　许多真核细胞的蛋白质，最显著的是某些哺乳动物的蛋白质比细菌的蛋白质大很多，在 *E. coli* 中不容易合成；
- **蛋白折叠和溶解度**　通常，过度表达导致**包含体**（inclusion body）的产生——不溶性的错误折叠蛋白质的聚集。包含体易于纯化，但表达的蛋白质通常可能仅在使用强烈的变性条件时才被溶解，并且主要问题是如何进行有效的体外折叠。

努力增加产量和溶解度经常与**融合蛋白**（fusion protein）的制备有关，其中目的蛋白与内源性蛋白结合。例如，麦芽糖结合蛋白，硫还原素，泛素等等。修饰蛋白表达载体也是常用的方法，因此添加**亲和性标记**（affinity tag），通过亲和层析有助于重组体的纯化。两个受欢迎的体系是：

- **GST-谷胱甘肽亲和性**　谷胱甘肽-S-转移酶（GST）是对其底物谷胱甘肽具有很高亲和性的小蛋白。GST-融合蛋白可以制备（图 5.21B）并通过与含有谷胱甘肽的柱选择性结合被纯化。
- **多聚组氨酸-镍离子亲和性**　某些氨基酸如组氨酸的侧链对某些金属离子有很高的亲和性。处于这种情况下的表达载体通常导致连续六个组氨酸残基的亲和性标记结合（图 5.23）。$(His)_6$ 标记的侧链选择性并有力地与镍离子结合，通过使用镍-腈酸基质的亲和层析有助于其纯化。

表达文库

质粒载体经常用于表达克隆，因为它们易于操作，并且如果目的是表达感兴趣的特殊基因，它们是首选工具。然而，有时目的是制备大量不同的重组体作为表达资源，即**表达文库**（expression library）。在这种情况下，使用修饰的 λ 噬菌体载体经常是有用的，因为筛查大量重组体比较容易。

表达文库通常是通过使用 λ 载体如 λgt11 和 λZAP，克隆来自靶组织的 cDNA 而构建。含有单个噬菌体感染的细菌集落的滤膜可以通过接触抗体被筛选。然后阳性反应菌可以繁殖以便分离 cDNA 克隆，并且被分离的 cDNA 克隆反过来可以用于筛查基因组文库以鉴定同源基因。

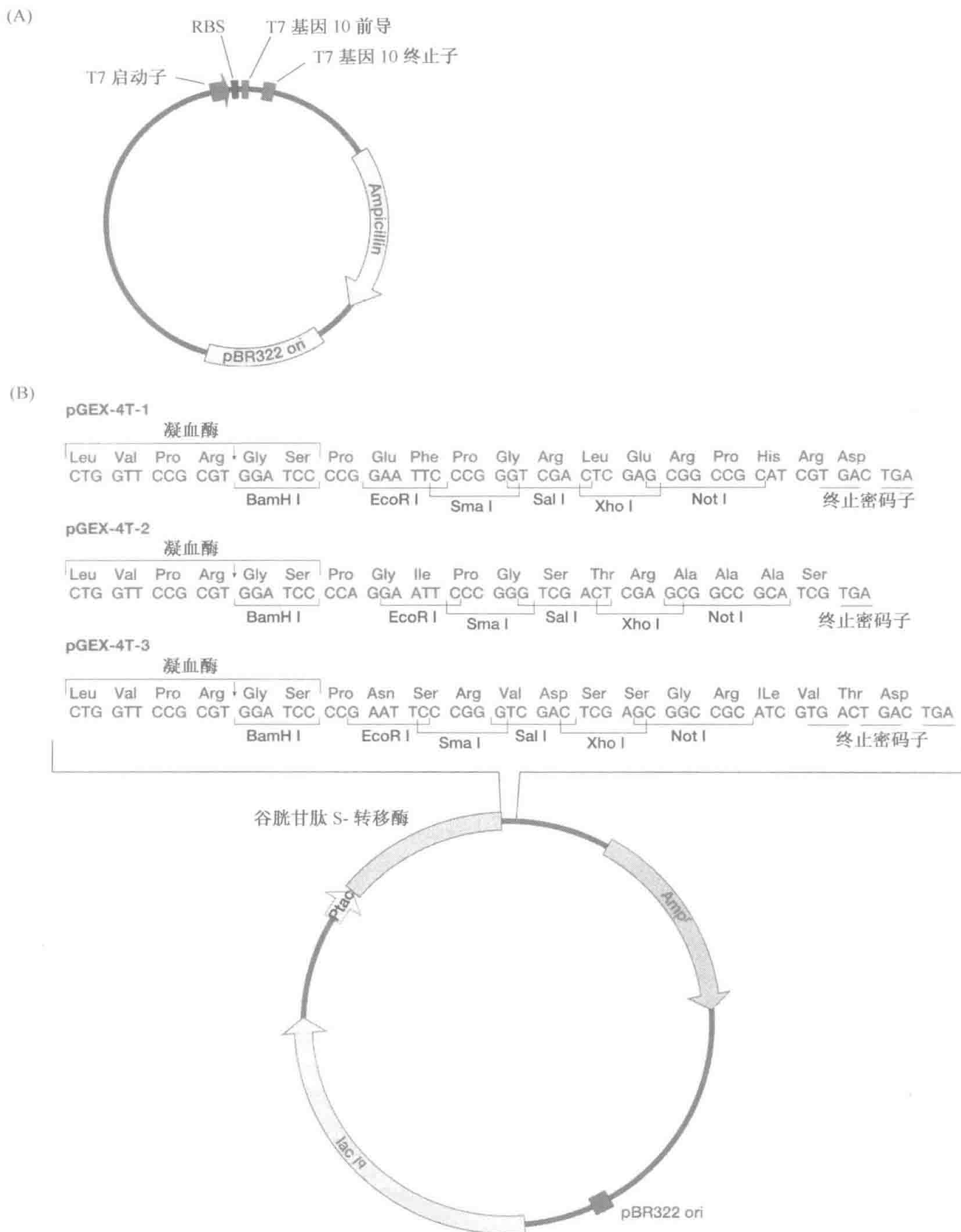


图 5.21　细菌表达载体的例子

(A) **pET-3 细菌表达载体**　利用噬菌体 T7 启动子有可能高水平表达。载体也含有编码一个 T7 基因 10 前导肽和一个 T7 基因 10 终止子（T7ter）的序列。在 *Nde*I 或 *Bam*H1 克隆位点进行克隆是可能的。如果是 *Bam*H1 克隆位点，制备的融合蛋白含有来自 T7 基因 10 的 13 个 N-端氨基酸，T7 基因 10 携带 T7 前导序列以确保高水平翻译。表达通过加入诱导位于被修饰的细菌染色体内的 T7RNA 聚合酶基因表达的 IPTG 而诱导。RBS，核糖体结合位点。(B) **pGET-4T 基因融合载体**　多克隆位点被定位，这样 tac 启动子引起转录后，制备融合蛋白，由 GST 的 N 端成分和目的蛋白的 C 端成分组成。开始于 *E. co*R1，pGEX-4T-1，pGEX-4T-2 和 pGEX-4T-3 的多克隆位点被排列，以至于所有三个氨基酸可读框是可能的。GST 融合蛋白在谷胱甘肽亲和纯化柱上易于纯化，如谷胱甘肽琼脂糖 4B，并且目的蛋白在凝血酶切割位点可以被切割。

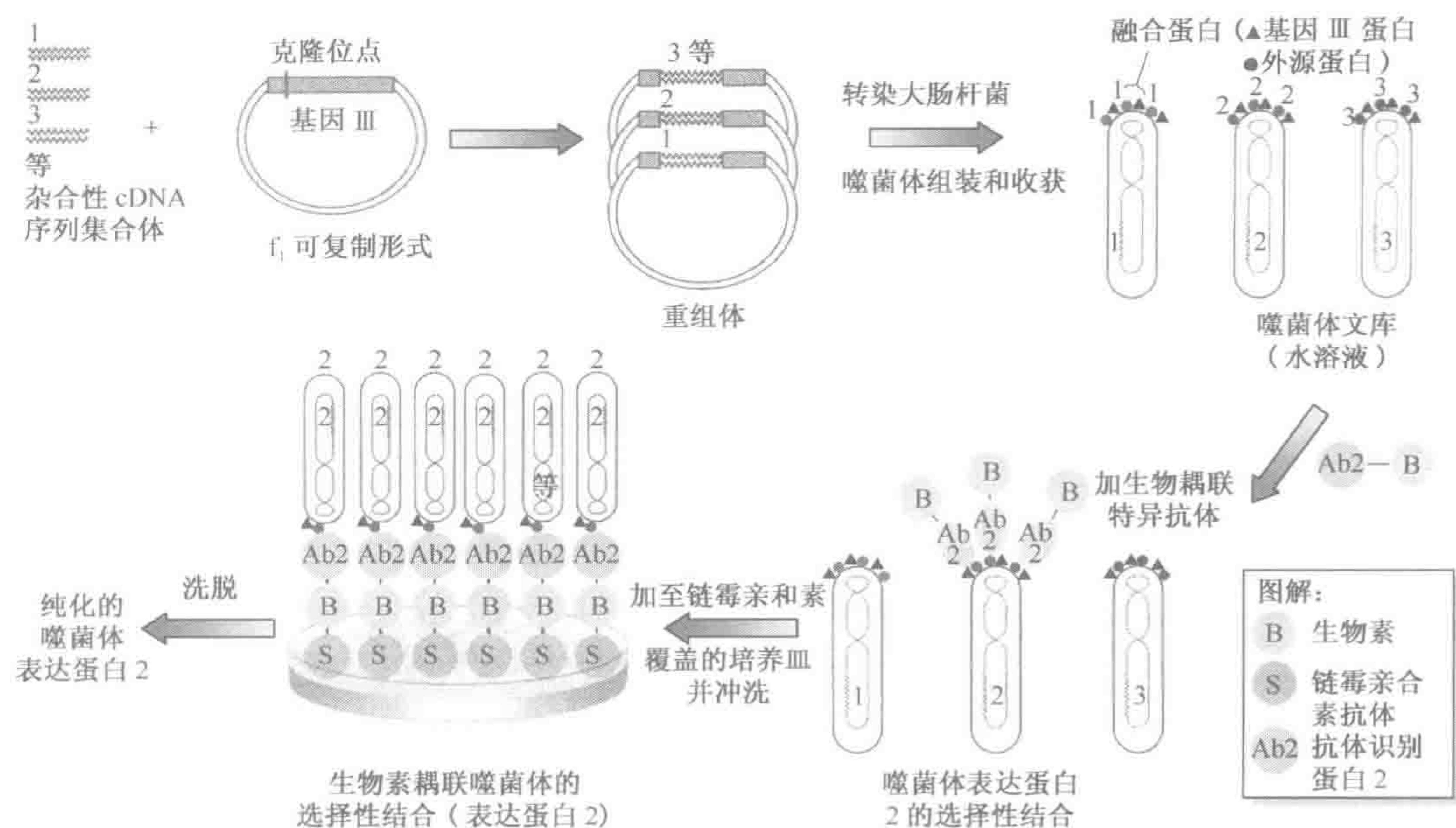


图 5.22　噬菌体展示

为了在噬菌体表面表达外源蛋白，外源 cDNA 被克隆进入噬菌体载体。在这里，DNA 插入编码微小噬菌体外壳蛋白的噬菌体 f1（或 M13，fd）的基因Ⅲ。克隆位点位于编码基因Ⅲ蛋白的最远 N 端序列的区域。噬菌体表达文库是通过 *E. coli* 转染，噬菌体组装，噬菌体从细胞中逸出并收获噬菌体而制备的（见图 5.18）。携带可读框内插入子的重组体经常可以表达，产生融合蛋白，其中 N 端成分由外源蛋白序列组成。与外源蛋白特异的抗体可以与展示序列的噬菌体特异结合，导致纯化。这样的**亲和纯化**（affinity purification）可以识别编码未知目的蛋白的 cDNA 序列（Parmley and Smith，1988）。

5.6.2　噬菌体展示是在细菌细胞表面表达蛋白质的一种表达克隆形式

噬菌体展示（phage display）是外源基因利用噬菌体的一种表达克隆的形式（Clackson and Wells，1994）。用遗传工程技术将外源 DNA 片段插入合适的噬菌体外壳蛋白基因。然后，被修饰的基因可以作为融合蛋白被表达，后者掺入病毒粒子并展示在噬菌体表面，然而，噬菌体保留感染性（**融合噬菌体**，fusion phage）。如果一种抗体对一种特殊蛋白质是可用的，噬菌体展示即蛋白质通过优先与抗体结合可以被筛选：在使用甚至微量的相关抗体时，含有靶决定子的病毒颗粒的亲和纯化可以从超过 10^8 倍量的不含决定子的噬菌体中完成。

起初，噬菌体展示与丝状噬菌体诸如 fd、f1、M13 的使用有关，其中外源基因掺入特定的次要外壳蛋白如基因Ⅲ蛋白中（图 5.23）。已经设计了几个有用的应用：

- **抗体工程**　噬菌体展示证明是构建抗体强有力的可选择的来源，包括人性化抗体，避开了免疫甚至杂交技术（Winter *et al.*，1994）；
- **一般蛋白质工程**　作为从突变文库筛查目的变异体的一种方法，噬菌体展示是随机诱变程序强有力的助手；
- **研究蛋白-蛋白相互作用**　这涉及基于文库的方法可用于鉴定与一指定蛋白质相互作

用的蛋白质。以同样的方式抗体可以应用于亲和筛查，一目的蛋白质（或蛋白能够结合的任何其他分子）被用作选择剂。该蛋白可以选择融合噬菌体，后者展示任何明显与其结合的其他蛋白质。

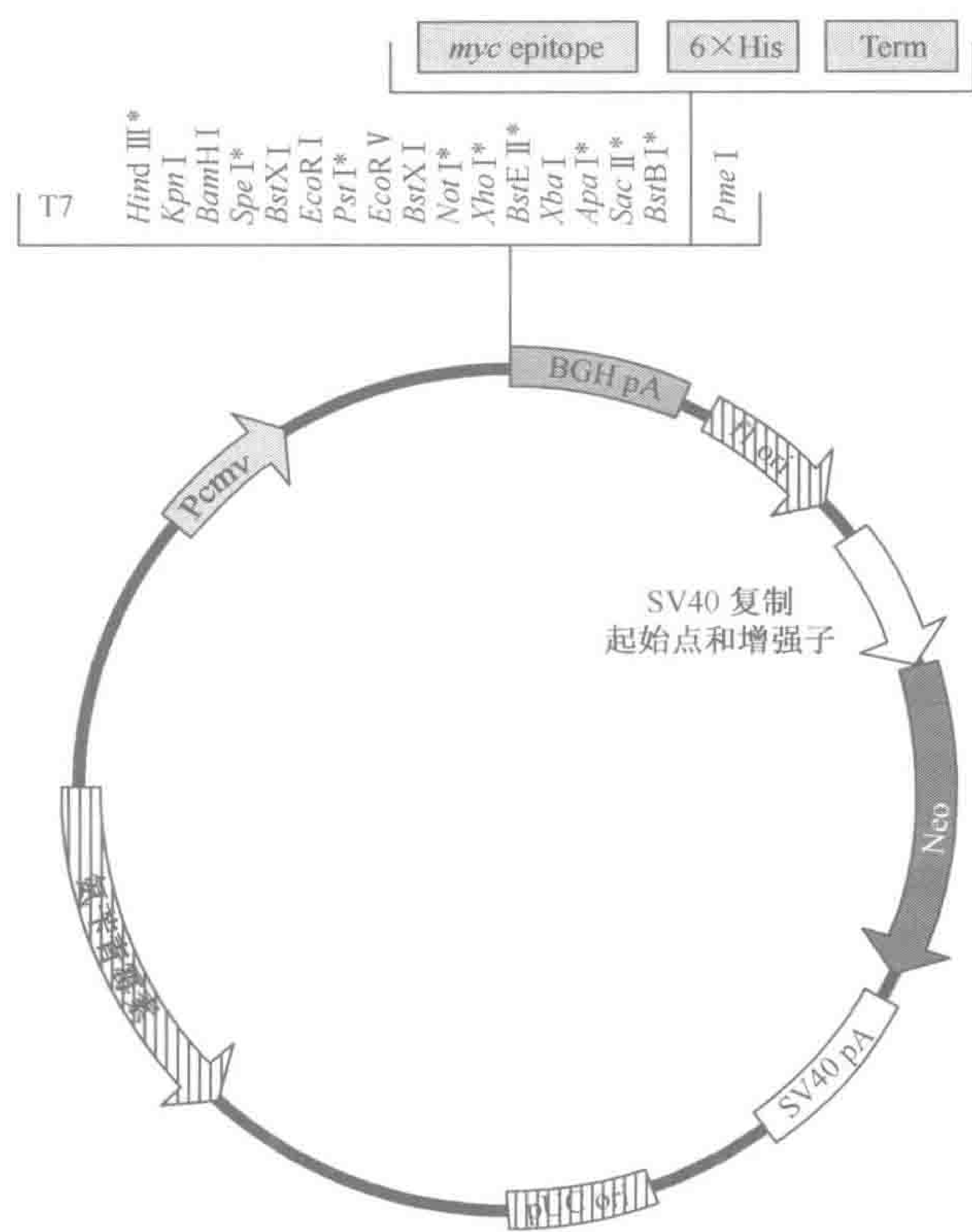


图 5.23　哺乳动物表达载体，pcDNA3.1/*myc*-HIS

Invitrogen 公司的质粒表达载体 pcDNA 系列在哺乳动物细胞提供高水平组成性表达。克隆的 cDNA 插入片段由于巨细胞病毒（P_{CMV}）的强大启动子（确保高水平表达）而被转录，该启动子携带牛生长激素多聚腺苷环化序列元件（BGHpA），可以使来自插入片段的 mRNA 产生限定的 3′端。*neo* 基因标记（由 SV40 启动子/增强子和 poly A 序列调节）通过在 G418 中的生长使筛选成为可能。顶部的多头连接包括多克隆位点、紧接着 6 个连续的组氨酸残基（6×His；使重组蛋白纯化易于进行）、一 *myc* 表位标签（允许利用与序列特异的抗体筛查重组蛋白）以及最终的翻译终止信号。用于 *E. coli* 繁殖的成分用▨表示，包括质粒 ColE1（pUCori）的复制起点、一种氨苄青霉素耐药基因（Amp）和一个 f1 起点，为制备单链重组体提供了选择（节 5.5.1）。

5.6.3　真核基因表达在真核细胞系中精确地进行

在细菌中合成的许多真核细胞的蛋白质的生物学特性由于缺乏正常的翻译后加工和不正确或无效的蛋白折叠可能不完全代表天然分子。当细菌表达体系具有表达高水平蛋白很强的优势时，其缺点已经促进真核细胞宿主包括昆虫和哺乳动物细胞对重组蛋白表达的有选择性使用。除了宿主的蛋白表达，哺乳动物细胞也经常作为一种方法，用于筛查体外操作转录和翻译后调控序列的作用。

使用动物的细胞系作为宿主细胞时，必须考虑外源 DNA 应该如何输入宿主细胞（框 5.5 和表 5.3）和表达的持续时间。在第二种情况下，可能有两种类型的表达体系：

▶ **瞬时表达**（transient expression）表达载体携带的 DNA 倾向于作为独立的遗传元件所谓的**附加体**（episome），保留在转染细胞内，而不是整合进入宿主细胞染色体。转基因的表达（已经导入动物或植物细胞的任何基因）在表达载体转染进入哺乳动物细胞系后约 2～3 天达到最高水平，但之后由于细胞死亡或表达载体的丧失，表达迅速减少；

框 5.5　转基因导入培养的动物细胞

各种方法可以用于转移基因进入动物细胞，但是分成两类：

▶ **转导**（transduction）　转导描述了病毒介导的基因转移。某些动物 DNA 和 RNA 病毒天然地感染人类和哺乳动物细胞。这些病毒的修饰使它们作为载体高效转移外源基因进入适当的靶细胞。它们提供以高拷贝数目表达的瞬时表达体系，正如腺病毒载体，也提供稳定表达，诸如依赖反转录病毒的那些一类 RNA 病毒。其自然生命周期与制备整合进入宿主细胞染色体的 cDNA 拷贝有关（表 5.3）。

▶**转染**（transfection）转染描述了非病毒介导的基因转移。注意转染一词与细菌中的转化过程相似。后者没有应用到转移基因进入动物细胞的过程，是由于它与改变的表型和未受限制的生长有关（见下文）。转基因可以通过不同方法转移：

- **使用非复制质粒载体**
- **使用携带病毒复制子的质粒载体。**通常使用 SV40 复制子（如含 COS 细胞的复制子；见 5.6.3），Epstein-Barr 病毒（人宿主细胞）或牛乳头状病毒（小鼠宿主细胞）。

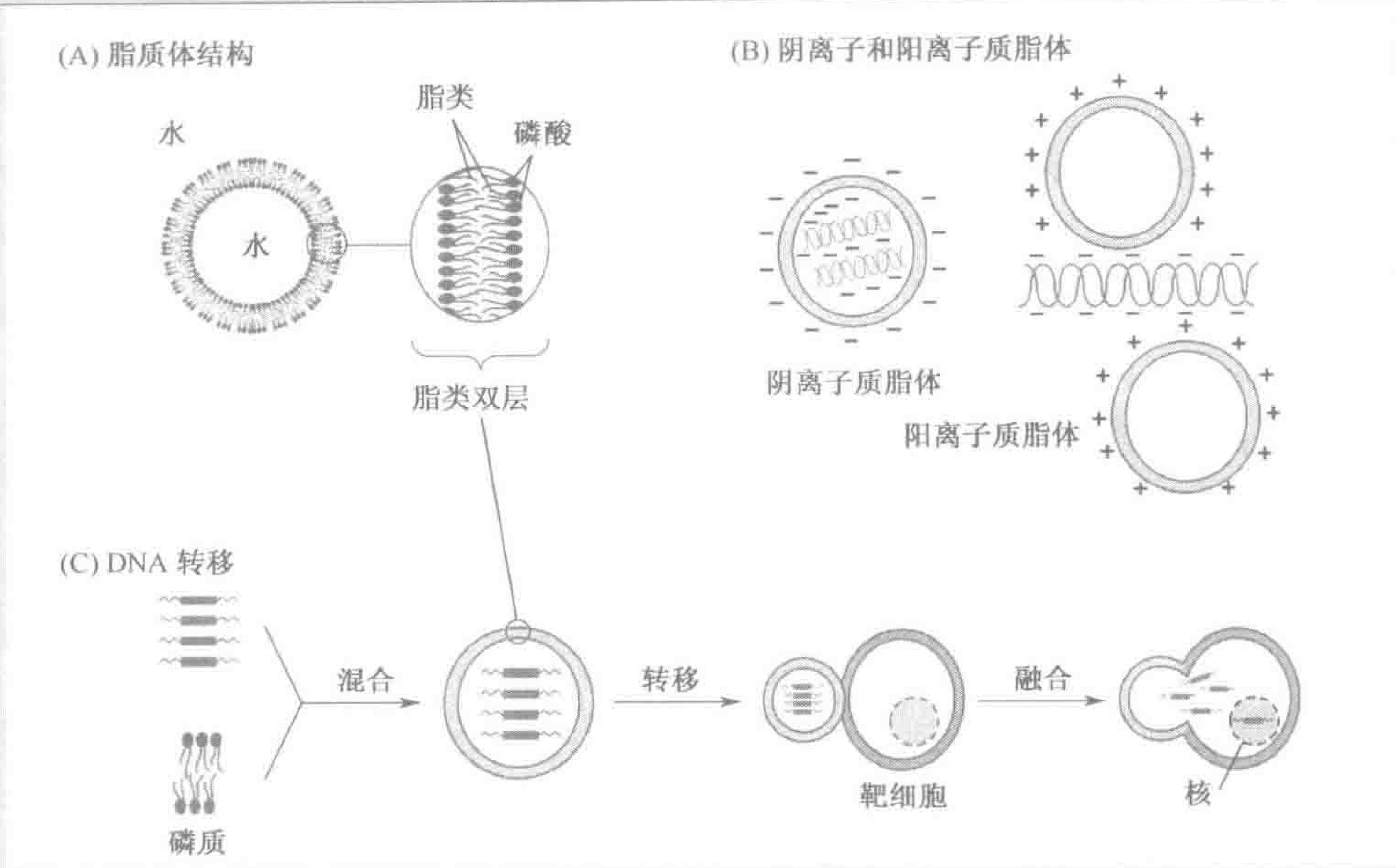
▶ 各种转染方法是可用的，包括使用：

- **磷酸钙**（calcium phosphate）　磷酸钙和 DNA 在靶细胞表面形成共沉淀。在质膜上高浓度的 DNA 可增加转染效率。
- **脂质体**（liposome）（人工脂质囊泡）　DNA 与其结合。脂质体在人工混合磷脂分子后，可以在水溶液中自发形成。脂质体可以与质膜融合，所以可以进入细胞内（见下文）；
- **电穿孔**（electroporation）　一种凭借电休克的流行方法用来在靶细胞内引起膜的暂时性去极化，借此有助于大的 DNA 分子穿过。

无论是通过转导还是转染，转移进入动物（或植物）细胞的基因被称为转基因（transgene）。转基因可以有如下不同的命运：

- **附加体转基因**（episomal transgene）　没有整合进入宿主细胞染色体的转基因可以以染色体外状态（附加体）存在于核内 。如果转基因与携带在宿主细胞中行使功能的复制起点的载体连接，它可以扩增，有时达到相当高的拷贝数。如果转基因没有与这样的复制起点连接，它在稀释（就像宿主细胞分裂一样）和降解之前只能维持短暂时间。
- **整合型转基因**（integrated transgene）　在某些情况下，转基因可以整合进入宿主细胞的染色体并稳定遗传。这种状态经常被描述为**稳定转化**（stable transformation）（注：转化一词的严格使用表示细胞表型被改变，这样细胞获得非限制生长特性）以及此形成的细胞被称为**细胞系**（cell line）。整合是效率极低的过程，所以，这罕见的稳定转化细胞必须借助筛选某种标记从非转化细胞的背景中分离（见本文）。

框 5.5　转基因导入培养的动物细胞（续）



▶ **稳定表达**（stable expression）表达载体携带的 DNA 设计整合进入宿主细胞染色体。这可能需耗时一个月来构建，但一旦构建，假设转基因能够表达，那么应该在所有细胞中发现表达产物。

表 5.3　在哺乳动物细胞中常见的病毒表达的载体体系

依赖性载体体系	宿主范围和位置	其他注释
腺病毒	广泛的哺乳动物宿主范围；常见的核附加体	插入大小仅至 8kb，高水平表达；高滴度重组病毒
腺病毒相关病毒	广泛的哺乳动物宿主范围；野生型病毒整合进入人 19 号染色体的特异位点	插入大小仅至 4.5Kb，需腺病毒包装；稳定表达
Epstein-Barr 病毒	作为核附加体存在于人、猴和狗，但啮齿类动物体不常见	
单纯性疱疹病毒	广泛的哺乳动物宿主范围；裂解	插入大小至 150Kb，通过删除相关病毒基因之一引起重组病毒复制缺陷
乳头状瘤病毒	作为核附加体存在于啮齿类（BPV）或人和猴（HPV）	用于研究基因调节和高水平转基因表达
多瘤病毒	广泛的哺乳动物宿主范围；可以整合	小鼠细胞中复制最佳；用于研究基因调节和高水平转基因表达
反转录病毒	可变的宿主范围但某些有广泛的哺乳动物宿主范围；作为 cDNA 拷贝整合进入宿主染色体	插入大小最大限制达 8.5Kb；低滴度的重组病毒；稳定表达
SV40	广泛哺乳动物宿主范围；可以整合但在 SV40 复制起点以及大 T 抗原存在下是附加型	
牛痘病毒	广泛哺乳动物宿主范围；裂解	主要用于转基因的过表达

使用杆状病毒在昆虫细胞中瞬时高水平蛋白表达

杆状病毒基因表达（baculovirus gene expression）是在昆虫宿主细胞制备大量重组蛋白的流行方法，蛋白产量比哺乳动物表达体系高，而花费低于哺乳动物表达体系。在大多数情况下，在昆虫细胞中表达的真核细胞蛋白的翻译后加工过程与发生在哺乳动物细胞中的蛋白加工过程相似，并且有可能表达非常大的蛋白质。结果，昆虫细胞中产生的蛋白质与哺乳动物细胞中表达的蛋白质有类似的生物学活性和免疫反应性。

通常用于蛋白表达的杆状病毒是加利福尼亚核多角体病毒（AcMNPV），可以在某些昆虫细胞系中繁殖。病毒多角体蛋白以高效率转录，尽管这对于病毒在其天然栖息地的繁殖是必要的，但在培养基中不需要。结果，其编码序列可以被编码外源蛋白的序列替代。克隆载体设计表达由强有力的多角体病毒启动子介导的异源蛋白，导致表达水平占细胞总蛋白的30%以上。

哺乳动物细胞中的瞬时表达

在哺乳动物细胞中表达哺乳动物蛋白质有明显的优势即正确的蛋白折叠和翻译后修饰没有问题并且有可能分析下游信号以及细胞效应。哺乳动物细胞中的稳定表达体系（典型的是建立在染色体整合的质粒序列的基础上）在产业化规模的**生物反应器**（bioreactor）中已经提供数千克的复杂蛋白质，但通常需要大量的时间、资源和设备投资。作为一种可选择的方法，大规模的瞬时表达体系已经开发用于在哺乳动物细胞中制备重组蛋白（Wurm and Bernard，1999）。

除了提供蛋白质表达，已经发现某些哺乳动物细胞系主要用于筛查体外操作对转录和转录后控制序列的作用。COS细胞提供了一个很好的例子，COS细胞是由SV40类人猿细胞系衍生的非洲绿猴肾细胞的稳定细胞系，CV-1。用SV40感染CV1细胞时，正常的SV40裂解周期发生。然而，Gluzman（1981）在含有突变复制起点的SV40基因组片段整合进入CV-1染色体后能够转化CV-1细胞。接着发生的COS（携有SV40复制起始点的CV1）细胞（携带SV40缺陷复制起点的CV1）在组成上（稳定地）表达SV40编码的**大T抗原**（large T antigen），它首先激活SV40复制起点需要的唯一病毒蛋白。通过以稳定方式表达SV40的大T抗原，COS细胞可以使任何导入的携带功能性SV40复制起点的环状DNA独立于宿主细胞染色体进行复制，而没有明确的大小限制。当瞬时表达载体转染进入COS细胞时，永生细胞系没有发生，因为大量的载体复制使细胞失活。即使仅有低比率的细胞成功地被转染，导入的DNA在那些细胞中以高拷贝数扩增可以补偿低的发生率。

哺乳动物细胞的稳定表达

转基因可以稳定整合进入宿主染色体DNA，但效率极低，所以罕见的稳定的转化细胞必须通过筛查某种标记从非转化细胞中分离。两个主要方法已经应用：

▶ **突变宿主细胞的功能性互补** 宿主细胞在某种程度上有遗传缺陷但原有功能可以通过**内生性的标记**（endogenous marker）而恢复。转基因和标记作为独立分子可以通过称为**共转化**（cotransformation）的过程而传递。一个例子是使用遗传学上有缺陷

的胸苷激酶（Tk^-）和 *Tk* 基因标记的细胞。TK 可以将胸苷转化为胸苷酸（TMP），但 TMP 也可以通过酶促反应利用 dUMP 合成。药物氨基喋呤阻断 dUMP→TMP 反应，所以有此药存在的细胞不能生长，除非细胞有胸苷来源和功能性的 *Tk* 基因。通常在 HAT（hypoxanthine，aminopterin，thymidine）培养基中筛选 Tk^+ 细胞；

▶ **可选择的显性标记的使用** 其主要缺陷是它们只能与突变细胞系一起使用，突变细胞系中相应的基因是没有功能的。结果是内生性标记已经在很大程度上被可选择的显性标记替代，后者赋予细胞全新的表型，因此可以用于任何类型细胞。这种类型的标记通常是赋予耐药性的细菌起源的药物耐受基因，已知药物影响真核和细菌细胞。例如氨基糖苷生素（包括新霉素和 G418）是细菌和真核细胞中蛋白合成的抑制剂。新霉素磷酸转移酶（neo）基因耐受新霉素。G418 等和用 *neo* 基因转化的细胞可以通过在 G418 中的生长被筛选。见图 5.23 携带 *neo* 标记物的哺乳动物表达载体的例子。

（付浩　译）

进一步阅读

Colosimo A, Goncz KK, Holmes AR *et al.* (2000) Transfer and expression of foreign genes in mammalian cells. *Biotechniques* **29**, 314–331.

Higgins SJ, Hames BD (1999) *Protein Expression. A Practical Approach.* Oxford University Press, Oxford.

Ling MM, Robinson BH (1997) Approaches to DNA mutagenesis: an overview. *Anal. Biochem.* **254**, 157–178.

McPherson MJ, Møller SG (2000) *PCR: The Basics.* BIOS Scientific Publishers, Oxford.

Old RW, Twyman RM, Primrose SB (2001) *Principles of Gene Manipulation*, 6th Edn. Blackwell Scientific Publications Ltd, Oxford.

REBASE database of restriction nucleases at http://rebase.neb.com/rebase/rebase.html

Sambrook J, Russell D (2001) *Molecular Cloning: A Laboratory Manual*, 3rd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

参考文献

Baneyx F (1999) Recombinant protein expression in *E. coli*. *Curr. Opin. Biotechnol.* **10**, 411–421.

Bedwell DM, Strobel SA, Yun K, Jongeward GD, Emr SD (1989) Sequence and structural requirements of a mitochondrial protein import signal defined by saturation cassette mutagenesis. *Mol. Cell. Biol.* **9**, 1014–1025.

Cheng S, Fockler C, Barnes WM, Higuchi R (1994) Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl Acad. Sci. USA* **91**, 5695–5699.

Clackson T, Wells JA (1994) In vitro selection from protein and peptide libraries. *Trends Biotechnol.* **12**, 173–184.

Cline J, Braman JC, Hogrefe HH (1996) PCR fidelity of Pfu DNA polymerase and other thermostable DNA polymerases. *Nucl. Acids Res.* **24**, 3546–3551.

Gluzman Y (1981) SV40-transformed simian cells support the replication of early SV40 mutants. *Cell* **23**, 175–182.

Higuchi R (1990) Recombinant PCR. In: *PCR Protocols. A Guide to Methods and Applications* (eds MA Innis, DH Gelfand, JJ Sninsky, TJ White). Academic Press, San Diego, pp. 177–183.

Iouannou PA, Amemiya CT, Games J, Kroisel PM, Shizuya H, Chen C, Batzer MA, de Jong P (1994) A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6**, 84–89.

Li HH, Gyllensten UB, Cui XF, Saiki RK, Erlich HA, Arnheim N (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**, 414–417.

Newton CR, Graham A. (1997) PCR. 2nd Edn. Springer-Verlag, New York.

Parmley SF, Smith GP (1998) Antibody-selectable filamentous fd phage vectors: affinity purification of target genes. *Gene* **73**, 305–318.

Schlessinger D (1990) Yeast artificial chromosomes: tools for mapping and analysis of complex genomes. *Trends Genet.* **6**, 248–258.

Sternberg N (1992) Cloning high molecular weight DNA fragments by the bacteriophage P1 system. *Trends Genet.* **8**, 11–16.

Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR (1994) Making antibodies by phage display technology. *Ann. Rev. Immunol.* **12**, 433–455.

Wurm F, Bernard A (1999) Large-scale transient expression in mammalian cells for recombinant protein production. *Curr. Opin. Biotechnol.* **10**, 156–159.

第6章　核酸杂交：原理和应用

本章内容

核酸杂交是分子遗传学的一个基本工具，它利用了单独的单链核酸分子形成双链分子的能力（也就是说：彼此**杂交**，hybridize）。为了发生杂交，相互作用的单链分子必须有充分高度的**碱基互补性**（base complementarity）。标准的核酸杂交实验涉及应用标记的核酸**探针**（probe）在一未标记的核酸分子-**靶**（target）核酸（**注**：当靶特指一个人想通过克隆扩增特异的 DNA 片段时，它在 DNA 克隆中有相当不同的用途）的复杂混合物中识别相关的 DNA 或 RNA 分子（也就是指具有明显高度序列相似性的分子）。

6.1　核酸探针的制备

在标准核酸杂交实验中，探针是以某种方式标记的。制备的核酸探针可以是单链分子，也可以是双链分子（图 6.1），但工作探针必须是单链形式。

常规的 **DNA 探针**（DNA probe）是通过以细胞为基础的 DNA 克隆或通过 PCR 分离的。在这两种情况下，开始时探针通常是双链的。细胞内克隆的 DNA 大小从 0.1kb 到数百 kb 不等，但通过 PCR 方法克隆的 DNA 长度通常小于 1kb。探针通常是在一个体外 DNA 合成反应中通过掺入标记的 dNTP 而标记的。

RNA 探针（RNA probe）来源于单链 RNA 分子，典型地为几百 bp 到几 kb 长。它们可以很方便地从已克隆至一专用质粒载体的 DNA 中获得，该载体在紧邻多克隆位点处有一噬菌体启动子序列。应用相关的噬菌体 RNA 聚合酶和四种 rNTP 可进行 RNA 合成反应，至少对其中的一种 rNTP 进行标记。随后就可以从克隆的插入片段产生特异性标记的 RNA 转录物。

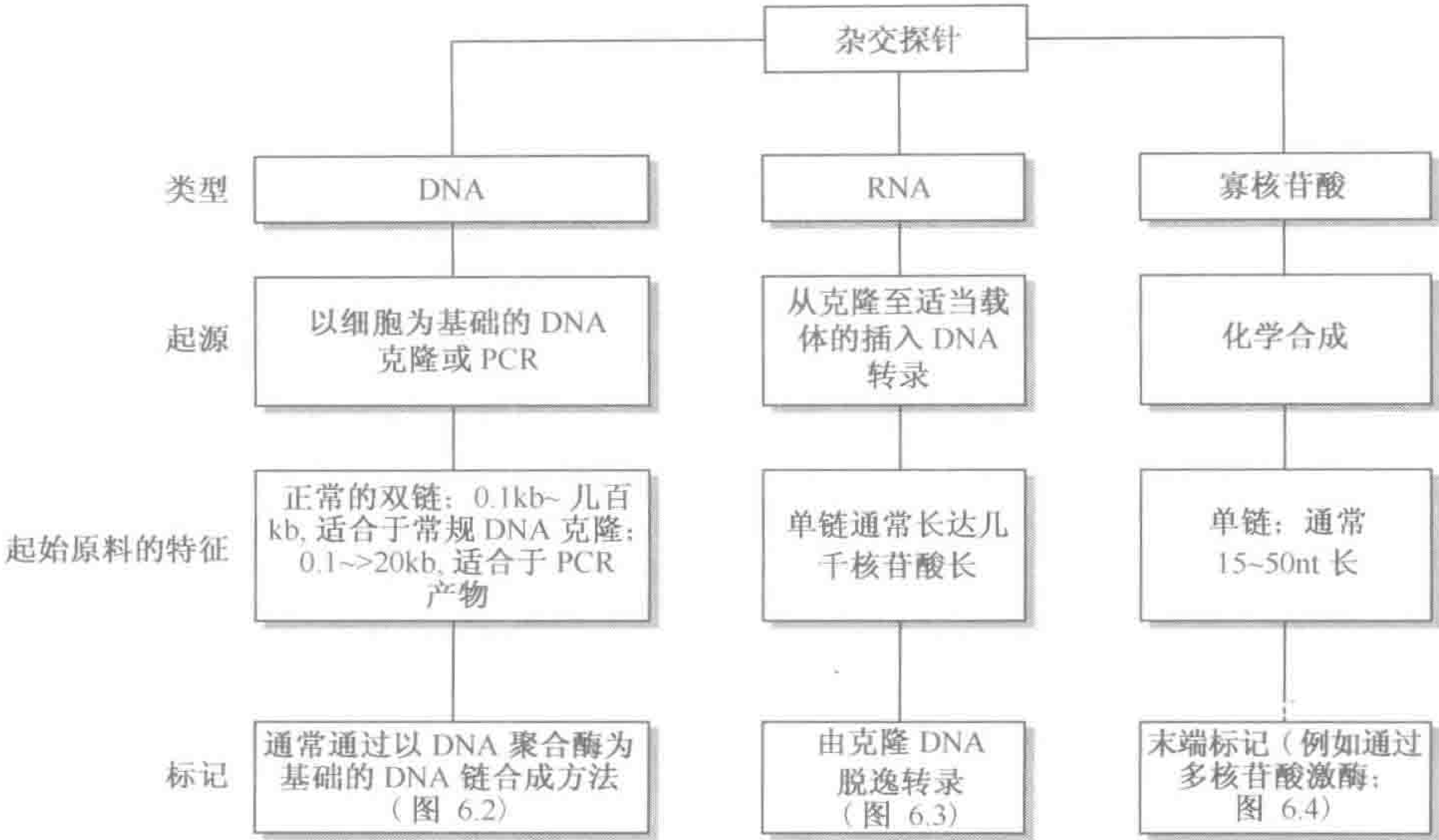


图 6.1　核酸杂交探针的起源及特性

寡核苷酸探针（oligonucleotide probe）是单链的并且非常短（典型地为 15～50 核苷酸长）。它们是通过化学合成的方法产生的（与来源于 DNA 克隆的其他探针不同）。将单核苷酸每次一个加至一个起始的单核苷酸上，一般是 3′端核苷酸最初固定于一固体支持物。通常，寡核苷酸探针是选择对靶 DNA 之前信息有反应的某一特异序列来设计的。然而，有时用**简并寡核苷酸**（degenerate oligonucleotide）作探针，包含一个平行合成的相关寡核苷酸集合的标记，这个探针一直设计为在特定的核苷酸位置相同，而应其他位置不同。寡核苷酸探针通常在 5′端用^{32}P 原子或其他标记基团标记。

6.1.1　通过修饰核苷酸的掺入能够在体外方便地标记核酸

通过向组织培养细胞中添加标记的脱氧核苷酸，能够在体内标记 DNA 和 RNA，但是这一过程限制了它的普遍应用。一种更加通用的方法是在**体外标记**（*in vitro* labeling）：通过应用一种适宜的酶来掺入标记的核苷酸，在体外标记纯化的 DNA、RNA 或寡核苷酸。广泛使用两种主要类型的方法：

- **链合成标记**（strand synthesis labeling）——标准的标记方法，在此方法中，应用 DNA 或 RNA 聚合酶来制作一起始 DNA 的 DNA 或 RNA 拷贝。在体外的 DNA 或 RNA 合成反应需要的四种核苷酸前体中至少有一种带有一标记基团。DNA 标记通常是通过三种方法之一进行的：缺口平移、随机引物标记及 PCR 介导的标记。RNA 标记是应用一个体外转录系统完成的。
- **末端标记**（end-labeling）——一种更加专有的标记方法，在此方法中仅仅在末端一个或几个核苷酸加入一标记基团。末端标记对于标记单链寡核苷酸（见下文）及**限制作图**（restriction mapping）是非常有用的。不可避免地由于仅掺入一个或很少的几个标记基团，因此标记的核酸的**特定活性**（specific activity）（总量除以掺入量标

记）比沿分子全长有多个标记核苷酸掺入的探针的特定活性低得多。

经缺口平移标记DNA

缺口平移（nick translation）过程包括在DNA中引入单链断裂（**缺口**，nick），留下暴露的3′羟基末端及5′磷酸末端。缺口可以通过加入一适宜的内切核酸酶如胰DNA酶Ⅰ获得。随后暴露的缺口可作为一个起始点，利用*E. coli* DNA聚合酶Ⅰ的DNA聚合酶活性在缺口的3′羟基侧引入新的核苷酸，同时，利用该酶的5′→3′外切核酸酶活性，从缺口的另一侧切除现存的核苷酸。结果缺口将沿着DNA（“被翻译的”）从5′→3′方向不断地移动（图6.2A）。如果反应是在一个相对低的温度（大约15℃）下进行，那么这个反应当完全更新现存的核苷酸序列时不再继续进行。尽管在这些温度没有最后的DNA合成，但是合成反应使得掺入标记的核苷酸代替先前存在的未标记的核苷酸。

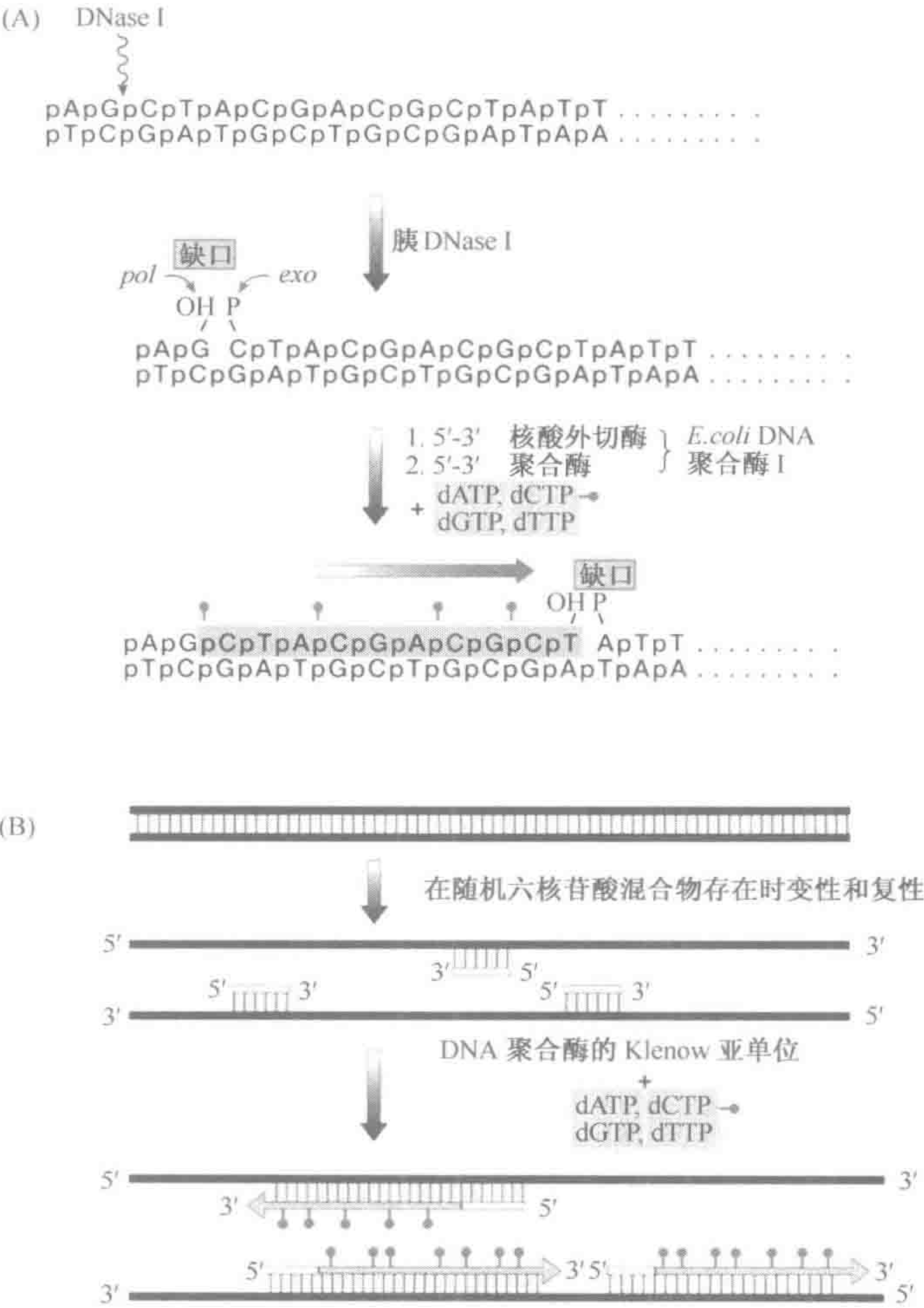


图6.2 通过体外DNA链合成进行DNA标记

(A) **缺口平移**（nick translation）。胰DNA酶Ⅰ通过切割内部磷酸二酯键（p）引入单链缺口，产生一个5′磷酸基团和一个3′羟基末端，加入的多亚单位酶*E. coli* DNA聚合酶Ⅰ具有两种酶活性：（Ⅰ）5′→3′外切核酸酶作用于缺口暴露的5′端，并按5′→3′方向连续地切除核苷酸。（Ⅱ）DNA聚合酶继续按5′→3′方向将新的核苷酸加至暴露的3′羟基上，从而替代由外切核酸酶切除的核苷酸，引起缺口的侧向移位（平移）。(B) **随机引物标记**（random primed labeling）。使用分离的DNA链作为模板，以及随机六核苷酸作为引物，*E. coli* DNA聚合酶Ⅰ的Klenow亚单位可以合成新的放射性标记的DNA链。

随机引物 DNA 标记

随机引物 DNA 标记方法 [有时也称为**寡标记法**（oligolabeling）；Feinberg and Vogelstein，1983] 是建立在一个所有可能的六核苷酸的混合物杂交基础之上：起始的 DNA 变性，然后缓慢冷却，以便各个六核苷酸能够与 DNA 链中适宜的互补序列结合。新互补 DNA 链的合成是以结合的六核苷酸为引物，由 *E. coli* DNA 聚合酶 I 的 Klenow 亚单位催化（具有聚合酶活性而无相关的 5′→3′外切核酸酶活性）。DNA 的合成在四种 dNTP 存在时进行，至少有一种含有一标记基团（图 6.2B）。这种方法产生具有高度特定活性的标记 DNA。因为所有的序列结合出现于六核苷酸混合物中，所以引物与模板 DNA 的结合以一种随机的方式发生，并且标记在全长 DNA 中是一致的。

PCR 介导的 DNA 标记

标准 PCR 反应能被修改为包括 1 个或多个标记的核苷酸前体，使其掺入至 PCR 产物，遍及其全长。

RNA 标记

RNA 探针（**核糖探针**，riboprobe）能够通过在体外转录的插入 DNA 克隆至一个具有噬菌体启动子的适宜质粒载体中而获得。例如，质粒载体 pSP64 在紧邻一多克隆位点处含有细菌噬菌体 SP6 的启动子序列。然后应用 SP6 RNA 聚合酶从 SP6 启动子序列的一特异起始点起始转录，任何插入到多克隆位点的 DNA 序列均可以转录。通过应用至少有一种 dNTP 被标记的 dNTP 混合物，就能够产生高度特定活性的放射性标记的转录物（图 6.3）。细菌噬菌体 T3 和 T7 启动子/RNA 聚合酶系统也常用于生产核糖探针。由任何克隆于这样载体中的基因均可产生标记的有义核糖探针和反义核糖探针（基因可按两个方向的任何一个克隆），并广泛应用于组织原位杂交（节 6.3.4）。

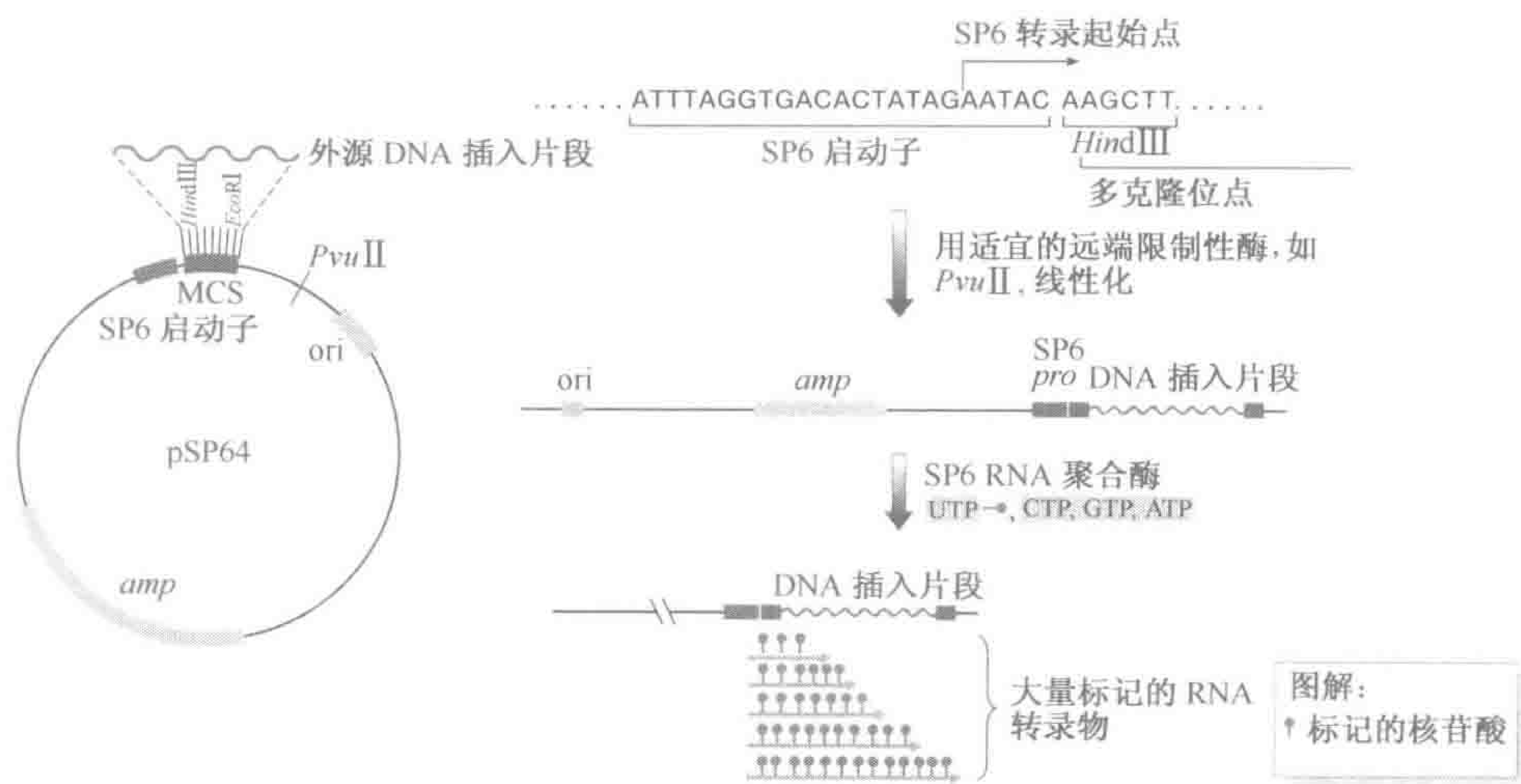


图 6.3 专门的质粒载体中克隆的 DNA 插入片段通过脱逸转录形成核糖探针（RNA 探针）
质粒载体 pSP64 除了具有复制起始位点（ori）和氨苄青霉素耐药基因（amp）外，在紧邻多克隆位点（MCS）处含有噬菌体 SP6 RNA 聚合酶的启动子序列。在 MCS 11 个单一限制酶切位点之一处克隆一个适宜的 DNA 片段后，在刚刚远离插入 DNA 的单一限制酶切位点用限制酶切割，使纯化的 DNA 重组线性化（本例中为 *Pvu* II）。此后应用 SP6 RNA 聚合酶和 NTP 混合物，能够产生标记的、插入片段特异性的 RNA 转录物，其中至少有一种被标记（本例为 UTP）。

末端标记

通常应用多核苷酸激酶（**激酶末端标记**，kinase end-labeling）对单链寡核苷酸标记。典型地，这种标记是以 ATPγ 磷酸位置^{32}P 的形式提供的，而多核苷酸催化其与 5′末端磷酸的交换反应（图 6.4A）。较大的 DNA 片段也能被末端标记，但通常通过可选择的方法，包括：

- **填充式末端标记**（fill-in end-labeling）（图 6.4B）——应用一适宜的限制酶处理 DNA，产生一个突出的 5′端，而聚合酶活性用于加入标记的互补的核苷酸以“填充”凹陷的末端。这通常应用 *E. coli* DNA 聚合酶 I 的 Klenow 亚单位（见上文）来完成。根据需要，标记的片段可以应用另一种限制性核酸酶在内部切割，从而产生两个在一个末端标记且能够按大小进行分离的片段。
- **引物介导的 5′端标记**（primer-mediated 5′ end-labeling）——一种简单的 PCR 方法，

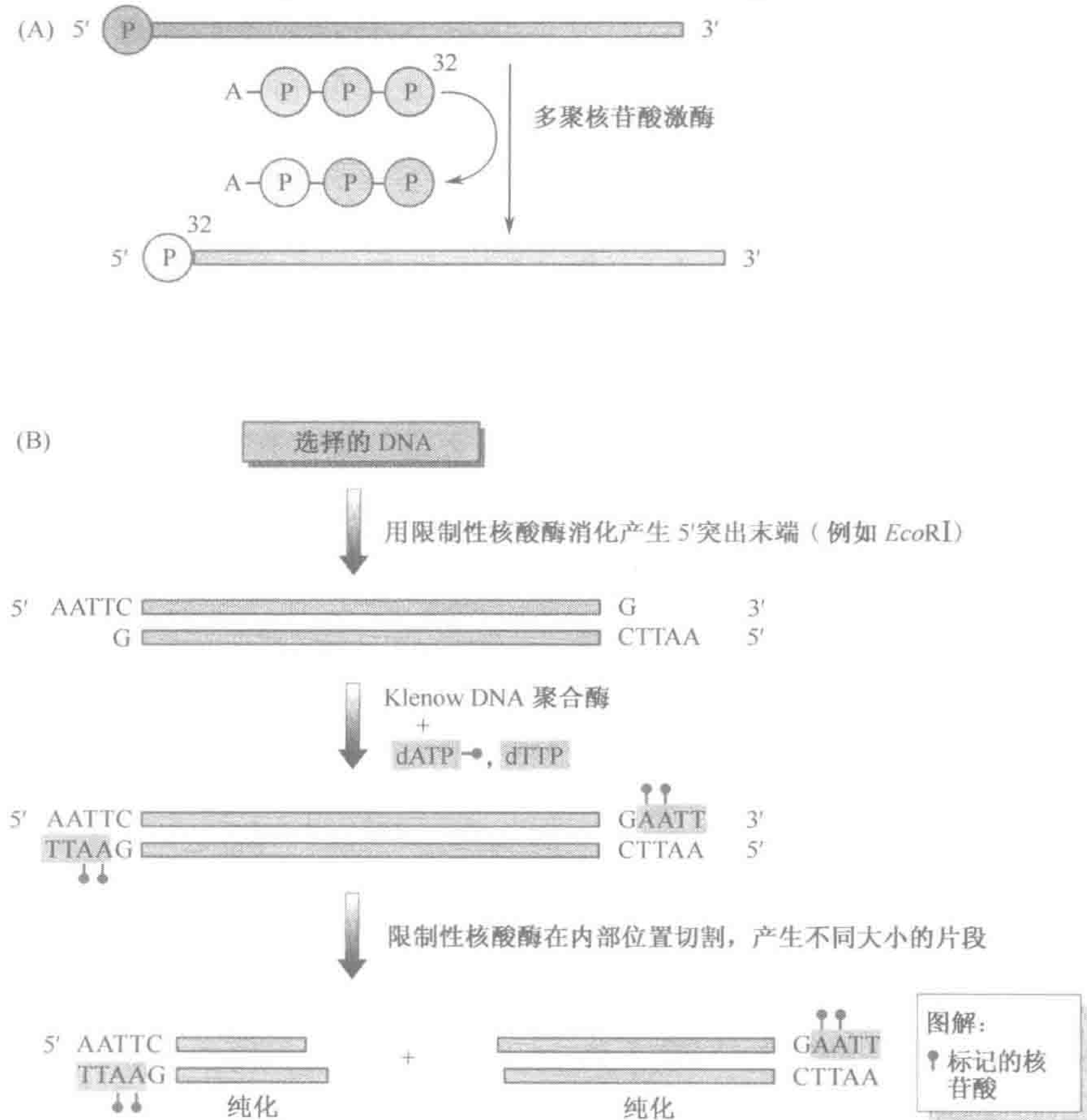


图 6.4　核酸末端标记

(A) **寡核苷酸的激酶末端标记**（kinase end-labeling of oligonucleotide）。寡核苷酸 5′端在交换反应中被［γ-^{32}P］ATP 的^{32}P 标记 γ 磷酸替代（γ-^{32}P）。这一过程也可被用于标记双链 DNA 的两个 5′端。(B) **填充式末端标记**（fill-in end-labeling）。用一适宜的限制性核酸酶切割目的 DNA，产生 5′突出末端，突出末端作为引物，在 Klenow DNA 聚合酶的作用下掺入与突出末端互补的标记核苷酸。仅在一端标记的片段能够通过某一适宜的限制性酶的内部切割形成，产生两个大小不同的、很容易按大小分离的片段。

标记基团结合于引物的5′端，随着PCR的进展，具有5′端标记的引物就掺入到PCR产物中。

6.1.2 通过同位素和非同位素方法能够标记核酸

同位素标记及检测

传统上，核酸是通过掺入含有放射性同位素的核苷酸而标记的。这样的**放射性标记探针**（radiolabeled probe）含有带有放射性同位素（通常是^{32}P、^{33}P、^{35}S或^{3}H）的核苷酸，能够在溶液或者更常见的在一固体标本中［**放射自显影**（autoradiography），框6.1］被特异性地检测。

框6.1 放射自显影原理

放射自显影（autoradiography）是在一固体样品内定位和记录一放射标记复合物的一种过程，涉及在照相感光乳剂内产生图像。在分子遗传学应用中，固体样品通常由按大小分离的DNA或蛋白质样品组成，它们包埋于一干胶内，固定于一干尼龙膜或硝酸纤维素滤膜表面，或位于载玻片上固定的染色质或组织样品中。照相感光乳剂由清澈的凝胶相的银卤化物晶体悬液组成。放射性核素释放的β粒子或γ射线穿过感光乳剂后Ag^{+}转换为Ag原子。随后一旦图像经过显影，结果产生的潜在的图像就转换成可视的图像。显影是全部银卤化物晶体还原产生金属银的放大过程。**固定**（fixing）过程造成任何未曝光的银卤化物晶体的清除，产生一个提供原始样品中放射标记分布的二维描述的放射自显影图像。

直接放射自显影（direct autoradiography）涉及紧贴X-射线胶片放置样品，胶片是涂有照相乳胶的塑料薄片。样品的放射性在显影的胶片上产生黑色区域。这种方法最适于检测弱到中等强度的β放射性核素（如^{3}H、^{35}S等），然而它不适用于高能量的β粒子（如来自^{32}P）；这样的发射物穿过胶片造成大多数能量的浪费。

间接放射自显影（indirect autoradiography）是一种修改的方法，通过一适宜的化学物品［闪烁体（scintillator）或荧光体（fluor）］将发射性能量转换成光。对于发射高能量辐射（如^{32}P）的样品来说，一种常用的方法是应用**加强屏**（intensifying screen）——一种固体无机闪烁体片，放置于胶片后方。那些穿过照相感光乳剂的发射物被屏吸收，并转换成光。通过有效地在直接放射自显影发射物上添加一照相发射物，图像被增强。

放射自显影信号的强度依赖于放射性同位素发出的辐射强度及曝光时间，曝光时间通常很长（1天或几天，在某些应用中甚至几周）。^{32}P已广泛应用于核酸杂交实验，因为它能发射高能量的β粒子，具有高度的检测敏感性。然而它的缺点在于其相对不稳定性（表6.1）。

表6.1 常用于标记DNA或RNA探针的放射性同位素的特点

放射性同位素	半衰期	衰变类型	放射能量
^{3}H	12.4年	β^{-}	0.019MeV
^{32}P	14.3天	β^{-}	1.710 MeV
^{33}P	25.5天	β^{-}	0.248 MeV
^{35}S	87.4天	β^{-}	0.167 MeV

^{32}P β 粒子发射的高能量在需要精确的物理分辨率来清楚地解释放射自显影图像的情况下可能是一种缺点。结果，在特定过程中更愿意选择放射较少能量 β 粒子的放射核素。例如^{35}S 和^{33}P 用于 DNA 测序和组织原位杂交，^{3}H 用于染色体原位杂交。^{35}S 和^{33}P具有中等半衰期，^{3}H 具有一个非常长的半衰期，但由于它放射的 β 粒子的较低的能量而变得不利，需要非常长的曝光时间。

用于 DNA 链合成标记反应中的^{32}P 标记和^{33}P 标记核苷酸在 α 磷酸位置标有放射性同位素，因为来自 dNTP 前体的 β 和 γ 磷酸并不掺入到不断增长的 DNA 链中。然而，激酶介导的末端标记应用 [γ-^{32}P] ATP（图 6.4A）。至于在 DNA 或 RNA 链合成过程中掺入的^{35}S-标记核苷酸，NTP 或 dNTP 在 α 磷酸基团的 O 位携带一个^{35}S 同位素。^{3}H 标记的核苷酸在几个位置携带放射性同位素。携带一个放射性同位素的分子的特异性检测通常通过放射自显影完成（框 6.1）。

非同位素标记及检测

非同位素标记系统涉及非放射性探针的使用，它广泛应用于许多不同领域（Kricka，1992）。介绍两种主要类型的非放射性标记：

▶ **直接非同位素标记**（direct nonisotopic labeling）——掺入一个含有附加标记基团的核苷酸。这一系统通常包括含有**荧光团**（fluorophore）的修饰核苷酸的掺入，这种荧光素是一个当暴露于某一特定波长光时能够发出荧光的化学基团（图 6.5 和框 6.2）。

框 6.2　荧光标记和检测系统

核酸的荧光标记是 20 世纪 80 年代发展起来的，并已证明在许多不同的应用领域具有极好的应用价值，包括染色体原位杂交、组织原位杂交及自动化 DNA 测序。荧光标记可通过掺入一个含有适宜**荧光基团**（fluorophore）（一种化学基团——当它暴露于某一特异波长的光时发出荧光）的修饰核苷酸（通常是 2′脱氧尿嘧啶 5′三磷酸）用于核酸的直接标记。用于直接标记的普通的荧光基团包括：**荧光素**（fluorescein）（一种淡绿色荧光染料）、**罗丹明**（rhodamine）（一种红色荧光染料）**和氨甲基香豆素**（amino methyl coumarin）（一种蓝色荧光染料）（图 6.5A）。可选择地，当含有一个报告基团（如生物素或地高辛）的修饰核苷酸被掺入至核酸时（图 6.6），可以应用间接标记系统，然后报告基团与亲和分子（如抗生物素蛋白链菌素或地高辛特异抗体）特异性结合，此亲和分子附带一荧光基团，如氨甲基香豆素乙酸（AMCA），荧光素异硫氰酸酯（FITC）或其他荧光素衍生物，以及四甲氨基罗丹明异硫氰酸酯（TRITC）及其他罗丹明衍生物。

直接或间接标记系统中荧光基团的检测是通过从适宜的光源（如荧光显微镜的汞蒸气灯，自动 DNA 测序仪的氩激光）发出一束光，通过适宜的，设计为在期望的激发波长处发射光的颜色滤光片（激发滤光片）而完成的。在荧光显微镜系统中，这种光通过分色镜（反射某种波长的光，而允许其他波长的光直接通过）被反射至显微镜载玻片的荧光标记样品上。然后这种光激发荧光基团发出荧光，而当其这样做时，它在一个较长的波长，即**发射波长**（emission wavelength）处发射光。荧光基团发出的光而后向上，直接通过分色镜，通过适宜的遮挡滤片，随后透射至显微镜的目镜，也可以使用适宜的 CCD（电荷耦联设备）照相机捕捉。常见荧光基团的最大发射及激发波长见下表：

荧光基团	颜色	最大激发波长（nm）	最大发射波长（nm）
AMCA	蓝色	399	446
荧光素	绿色	494	523
CY3	红色	552	565
罗丹明	红色	555	580
Texas Red	红色	590	615

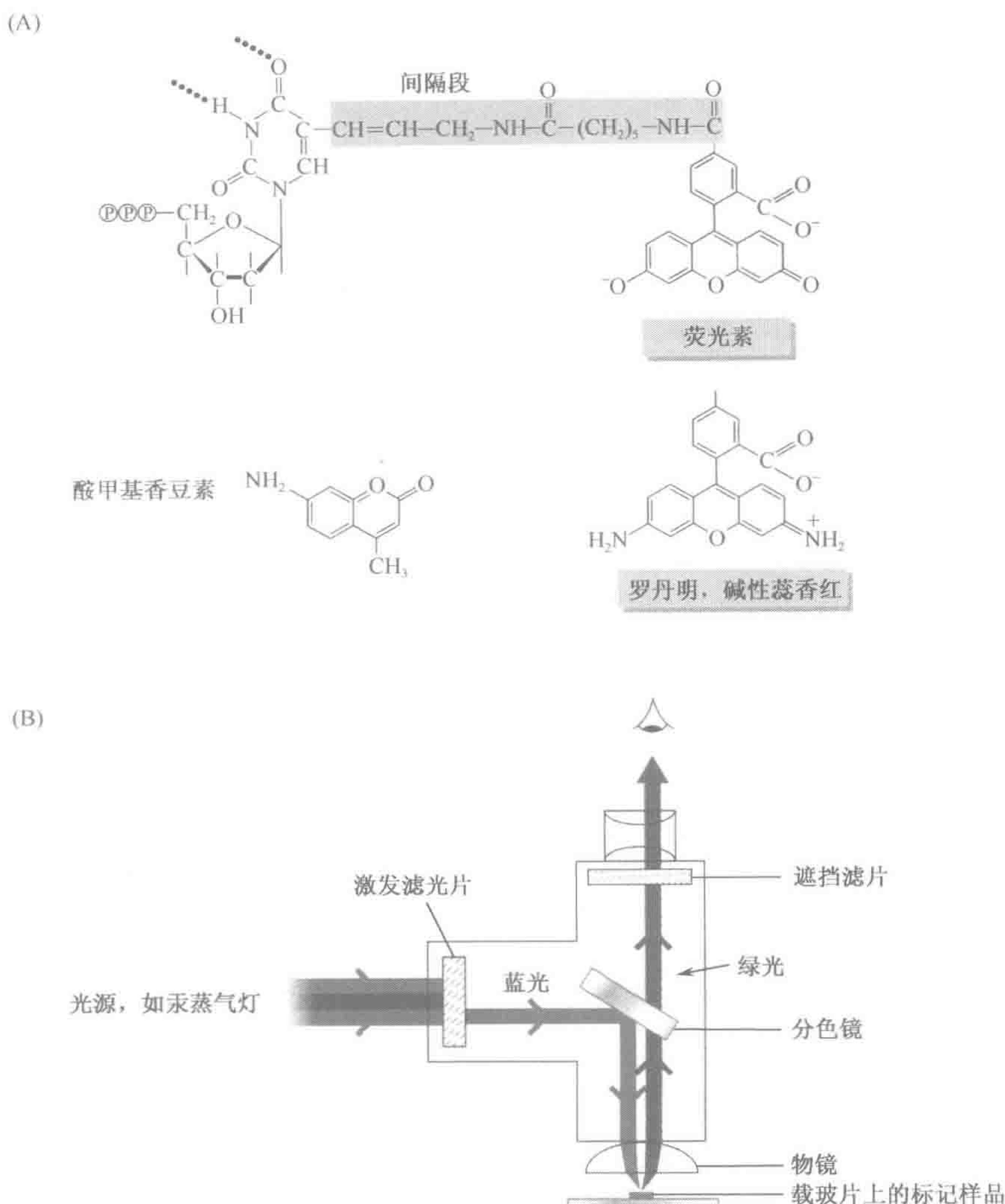


图 6.5　荧光显微镜和常用荧光素的结构

(A) **荧光素的结构**：上面的例子显示荧光素-dUTP。荧光素基团通过一个间隔段基团与尿嘧啶的 5′碳原子相连，结果当修饰的核苷酸被掺入至 DNA 中时，荧光素基团也很容易被掺入进去。下面是罗丹明的结构，很多荧光素基团衍生于该物质。(B) **荧光显微镜**：激发滤光片是一种颜色遮挡滤光片，在本例中仅允许蓝光通过。发射的蓝光具有一个适宜的波长，可被分色镜（光束分解）反射至标记的样品上，随后发出荧光或较长波长的光，本例中为绿光。发射的绿光的较长波长意味着它可直接通过分色镜。光随后通过第二个颜色遮挡滤光片，此颜色遮挡滤光片阻挡不想要的荧光信号而留下希望获得的绿色荧光发射物，使其通过并达到显微镜目镜。第二个光束分解设备也可允许光被 CCD 照相机记录。

▶ **间接非同位素标记**（indirect nonisotopic labeling）通常以化学偶联一修饰**报告分子**（reporter molecule）至核苷酸前体为特色。该报告基团掺入 DNA 后能够与一个**亲和分子**（affinity molecule）特异性地结合，该亲和分子是一个与报告基团具有很高的亲和性的蛋白质或其他配体。能通过适宜的方法检测到的一个**标记**（marker）分子或基团与亲和分子结合（图 6.6）。修饰核苷酸上的报告分子需从核酸骨架上伸出足

够远，以便易于通过亲和分子对它们的检测，因此需要一长的碳原子**间隔距离**（spacer），将核苷酸与报告基团分开。

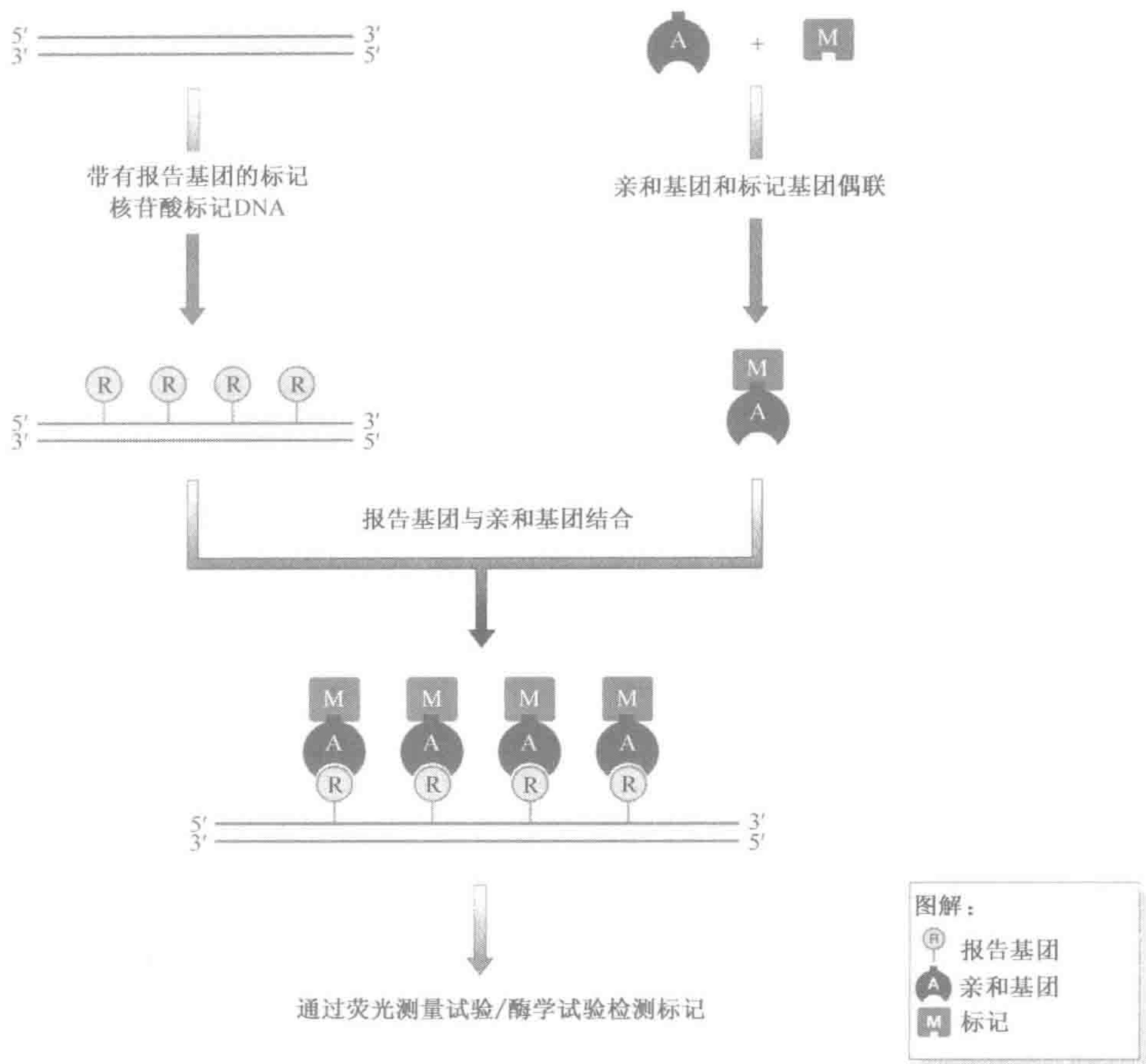


图 6.6　间接非同位素标记的一般原理

识别报告基团的蛋白质常是一个特异性抗体，如在地高辛系统中，或者与一特异性基团有很高亲和力的任何其他配体，如使用生物素作为报告子时的抗生物素蛋白链菌素（图 6.7）。可应用各种方法检测标记。如果它携带特异性荧光染料，可通过荧光测量实验检测。或者，它也可以是一个酶，如碱性磷酸酶，能够与一个酶学实验结合，产生一个能够通过比色法实验测量的产物。

两种间接非同位素标记系统被广泛应用：

- **生物素—抗生物素蛋白链菌素**（biotin-streptavidin）系统利用了两种配体极高的亲和性：以报告子起作用的**生物素**（biotin）（一种天然存在的维生素）和作为亲和分子的细菌蛋白**抗生物素蛋白链菌素**（streptavidin）。生物素和抗生物素蛋白链菌素非常牢固地结合在一起，亲和系数为 10^{-14}，是生物学已知的最强系数之一。生物素酰化探针能够很容易地通过在标记反应中加入适宜的生物素酰化核苷酸制备（图 6.7）。
- **地高辛**（digoxigenin）是一种植物类固醇（获自 *Digitalis* 植物），它的特异性抗体已经问世。地高辛特异性抗体可以检测掺入含有地高辛报告基团核苷酸的核酸分子（图 6.7）。

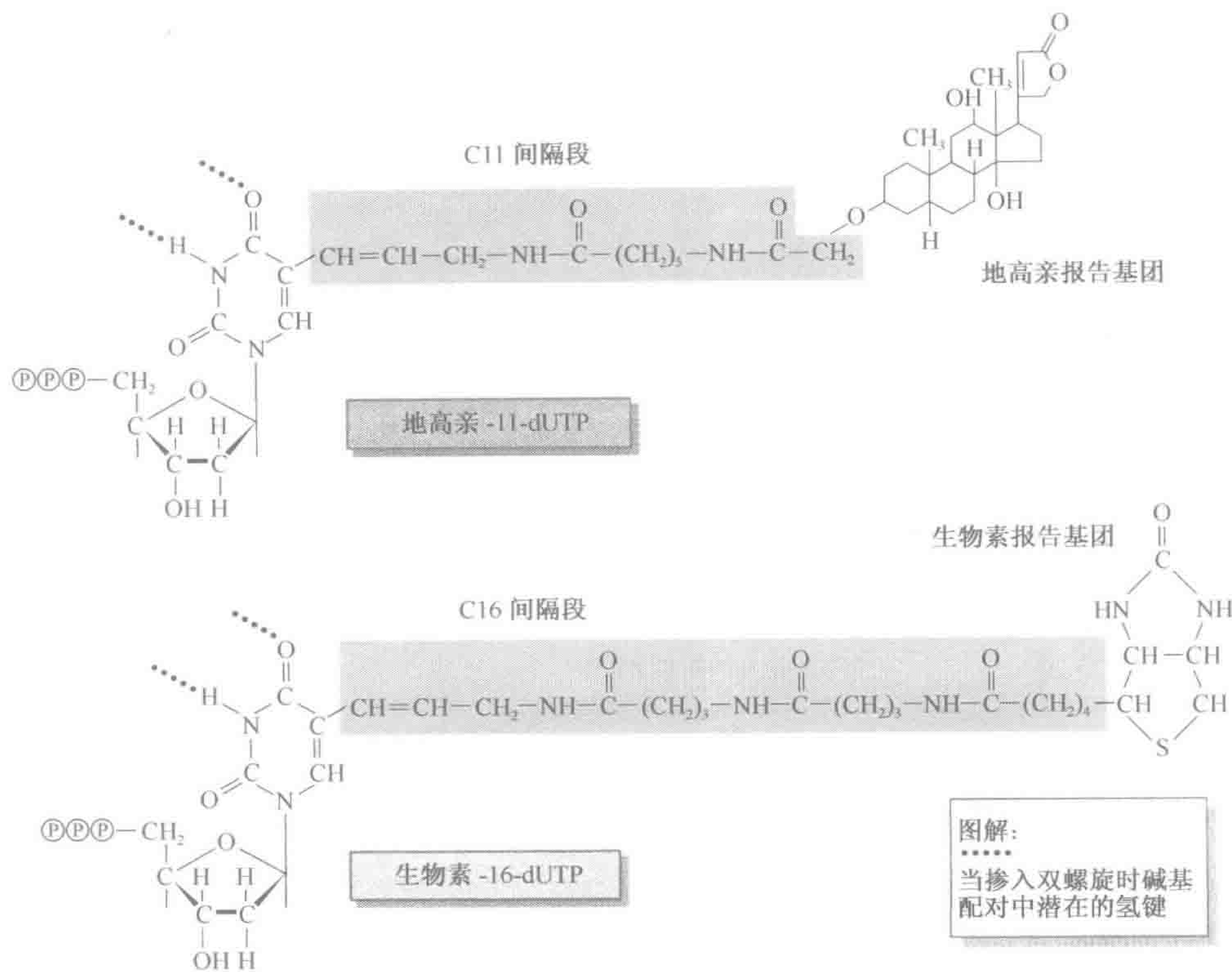


图 6.7　地高辛修饰和生物素修饰核苷酸的结构

注意在这些例子中，地高辛和生物素基团分别通过由总计 11 个碳原子（地高辛-11-UTP）或 16 个碳原子（生物素-16-UTP）组成的间隔段基团与 dUTP 尿嘧啶的 5′碳原子相连。地高辛和生物素基团是**报告基团**（reporter group），在掺入某个核酸后，它们与含有一个时常标记如荧光基团的特定配体结合。

各种不同的标记基团或分子能够与亲和分子如抗生物素蛋白链菌素或地高辛特异性抗体偶联。它们包括各种**荧光团**（fluorophore）（框 6.2）或酶，诸如碱性磷酸酶和过氧化物酶，可以通过比色法实验或化学发光实验等检测。

6.2　核酸杂交原理

核酸杂交是分子遗传学的一项基本技术。相关术语的词汇表见框 6.3，为不熟悉这些术语的读者提供帮助。

6.2.1　核酸杂交是在两个核酸群体中识别紧密相关分子的一种方法

定义和基本原理

核酸杂交的常用目的是获取一些关于一个不完全了解及非常复杂的核酸群体（靶核酸）的信息。这通过使用已知的核酸分子群体作为探针去识别靶核酸中紧密相关的核酸而实现。

探针和靶核酸之间相互作用的特异性来源于**碱基互补性**（base complementarity），因为两个群体均应用此方式处理，以确保存在的所有核酸序列都为单链。因此，如果探针或靶核酸起始是双链的，那么单个的链必须被分开（**变性**，denatured），一般通过加热或者碱处理。探针的单链和靶核酸的单链混合后，具有互补碱基序列的链允许**重新联合**（reassociate）（**复性**，reanneal）形成双链核酸。这一过程发生时形成两种类型的产物：

- **同源双链**（homoduplex）——探针或靶核酸内的互补链复性，重新产生最初发现于探针或靶群体中的双链分子。
- **异源双链**（heteroduplex）——是探针内一条单链核酸与靶群体内一条互补链碱基配对。这里所形成的双链核酸是一个新的核酸链组合。异源双链解释了一个核酸杂交实验的用途，因为杂交实验的全部目的是利用已知的探针识别靶核酸中相关的核酸片段（图 6.8）。

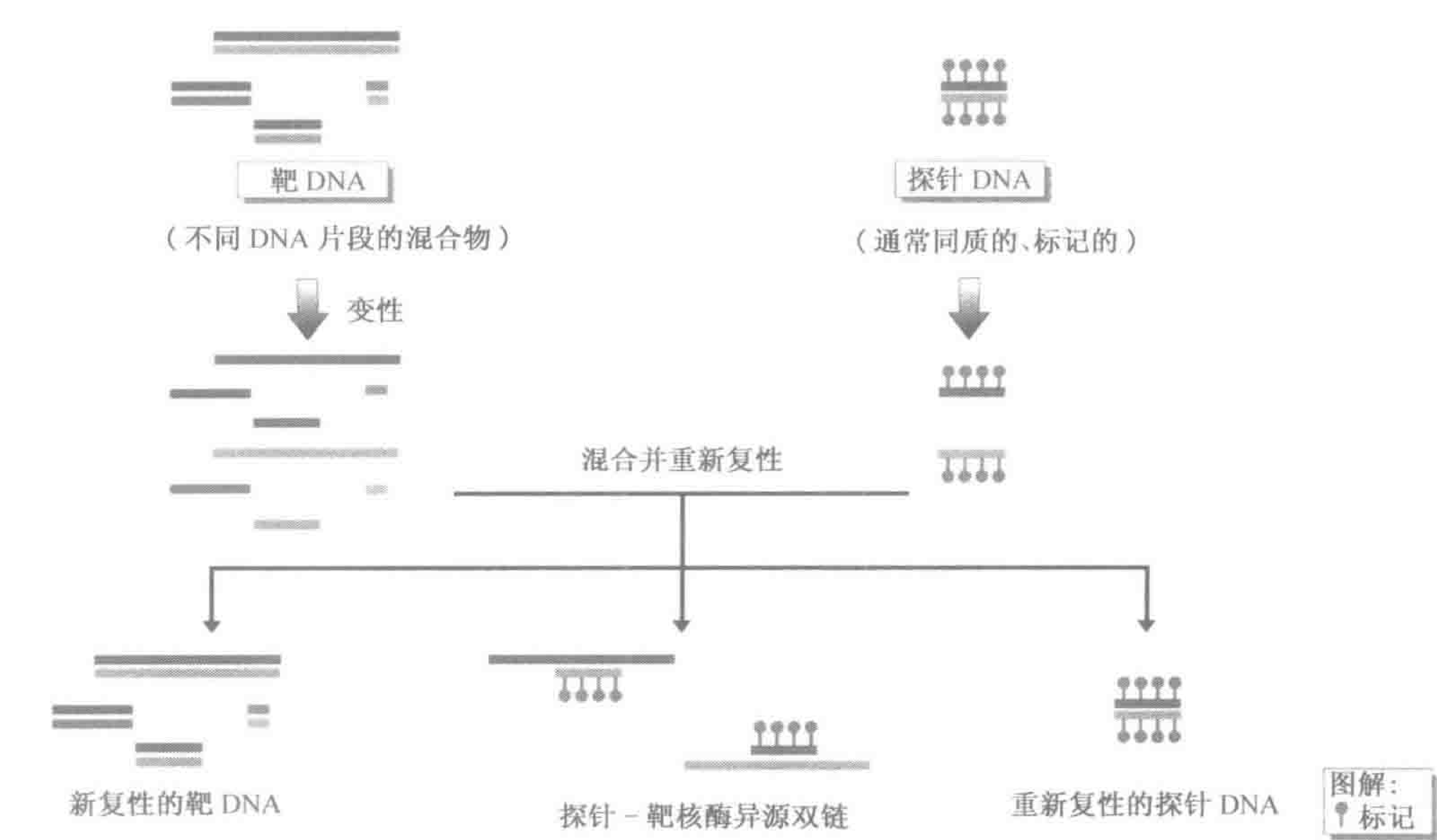


图 6.8　一个核酸杂交实验需要标记的单链核酸探针与一个靶核酸内互补序列间异源双链的形成
设想探针与靶分子的许多类型核酸分子中的一种自中心部分在序列上强相关。变性探针与变性靶分子的混合将会造成重新退火的探针-探针同源双链（底部右侧）和靶-靶同源双链（底部左侧），但也会造成在探针 DNA 和任何在序列上有明显相关的靶 DNA 分子间形成的异源双链（底部中间）。如果一种方法可用于移走未与靶 DNA 结合的探针 DNA，那么异源双链通过能够检测标记的方法很容易地被识别。

探针通常是同质的核酸群体（如特异克隆的 DNA 或化学合成的寡核苷酸），它们常被标记且位于溶液中。相反，靶核酸群体是典型地未标记的复杂群体，常结合于一固相支持物。然而，一些重要的杂交实验使用结合于一个固相支持物的未标记探针，用于寻找溶液中标记的靶 DNA 群。

熔解温度和杂交严格性

双链探针 DNA 变性一般通过加热一个标记 DNA 的溶液至某一温度，该温度足以破坏维持两条互补 DNA 链在一起的氢键。分开两条完全互补 DNA 链所需的能量依赖于许多因素，主要有：

▶ **链的长度**（strand length）——长的同源双链含有大量氢键，需要更多的能量来分开它们：由于标记过程代表性地产生短的 DNA 探针，所以这种影响对于 500bp 的初始（即标记前）长度是可以忽略的。

▶ **碱基成分**（base composition）——GC 碱基对比 AT 碱基对多一个氢键，因此 GC 含量高的链比 GC 含量低的链更难分开。

▶ **化学环境**（chemical environment）——单价阳离子（如 Na^+ 离子）的存在能稳定双链，而一些强极性分子，如**甲酰胺**（H-CO-NH_3^+）和**尿素**（H-$_3N^+$-CO-NH_3^+）作为化学变性剂发挥作用：它们通过破坏碱基对间的氢键使双链不稳定。

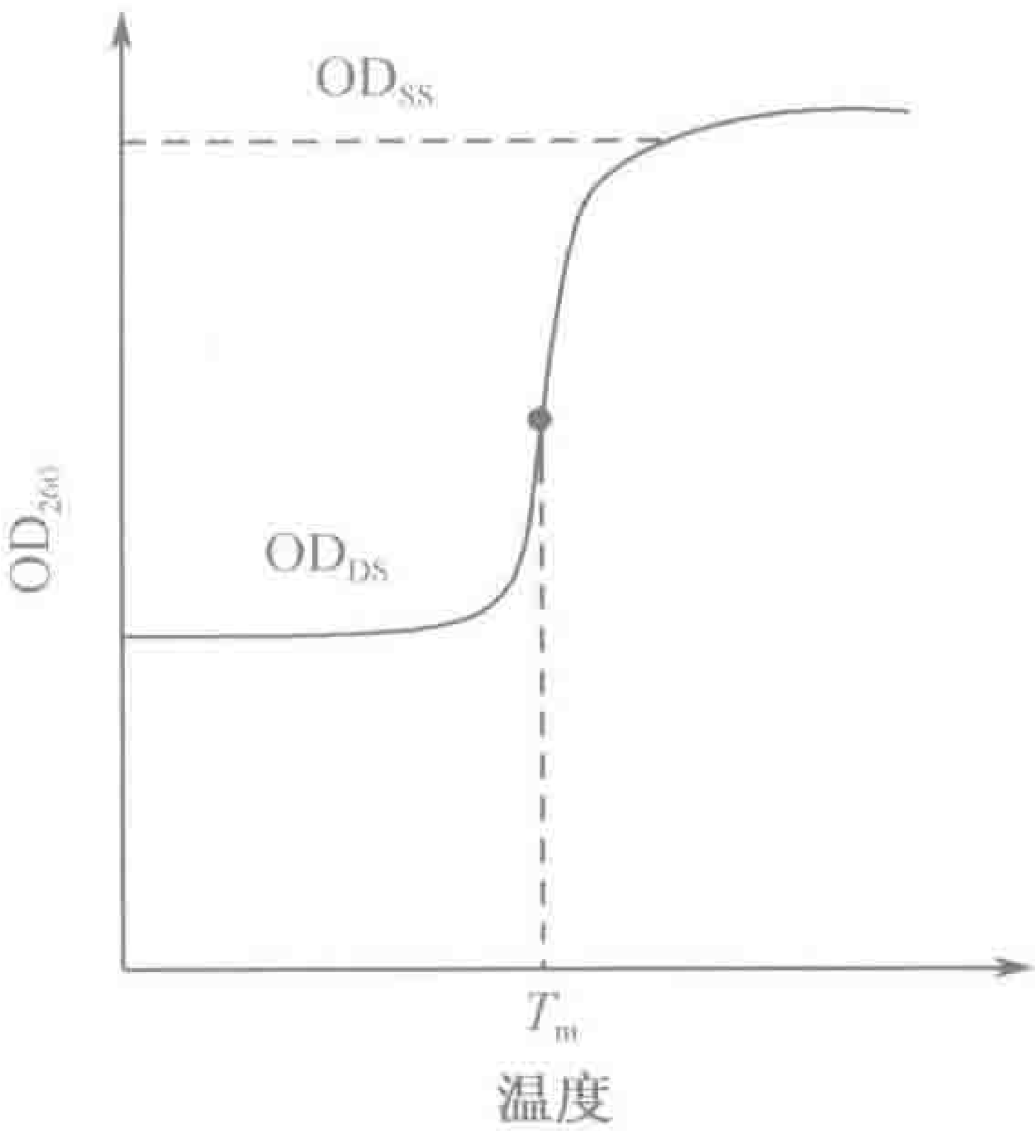


图 6.9　DNA 变性造成光密度增高　OD_{SS}和 OD_{DS}分别表示单链和双链 DNA 的光密度。它们之间的差异代表减色效应（见正文）。

一种有用的核酸双链稳定性的衡量标准是**熔解温度**（melting temperature，T_m）。这是与观察到的从双链形式转换成单链形式的中点温度一致的温度。这种转换可方便地通过 DNA 的**光密度值**（optical density，OD）来完成。核酸碱基强烈地吸收 260nm 紫外光。然而，双链 DNA 吸收的光比游离核苷酸吸收的光少得多。这种差异，即所谓的**减色效应**（hypochromic effect），是由于相邻碱基电荷系统间的相互作用造成的，这种相互作用是由双螺旋中相邻碱基平行堆积的方式产生的。因此，如果双链 DNA 被逐渐加热，则在 260nm 处吸收光增加（OD_{260}）接近于游离碱基的特征值。在光密度转变的中间值看作为 T_m 值（图 6.9）。

表 6.2　计算 T_m 值的公式

杂交体	T_m(℃)
DNA-DNA	$81.5+16.6(\log_{10}[Na^+]^a)+0.41(\%GC^b)-500/L^c$
DNA-RNA 或 RNA-RNA	$79.8+18.5(\log_{10}[Na^+]^a)+0.58(\%GC^b)+11.8(\%GC^b)^2-820/L^c$
oligo-DNA 或 oligo-RNA[d]：	
对于<20 个核苷酸	$2(l_n)$
对于 20～35 个核苷酸	$22+1.46(l_n)$

a 或者用于其他一价阳离子，但仅在 0.01～0.4M 范围内精确

b 仅精确于 30%～75% GC

c L＝双链碱基对的长度

d oligo＝寡核苷酸；l_n＝引物的有效长度，＝2×(G+C 数量)+(A+T 数量)

注：每 1%的甲酰胺，T_m 值降低大约 0.6 ℃，而 6M 尿素的存在降低 T_m 值大约 30 ℃。

对于哺乳动物基因组来说，具有大约 40%GC 的碱基组成时，在接近生理条件下，T_m 值大约为 87℃时 DNA 变性。DNA、RNA 或寡核苷酸探针形成正确杂交体的 T_m 可根据表 6.2 的公式计算。通常情况下，所选择的杂交条件是为了促进异源双链的形成，因此杂交温度通常较 T_m 低 25 ℃。然而杂交并清除过量的探针后，杂交冲洗要在更严格的条件下进行，为的是破坏除了密切相关序列间形成的双链之外的所有双链。

当双链形成区域含有完全的碱基匹配时，探针-靶序列异源双链在热力学上最稳定。一异源双链的两条链之间的错配降低 T_m：对于正常 DNA 探针来说，每 1%错配大致降低 T_m 值 1 ℃。尽管探针-靶序列异源双链通常不像再复性探针同源双链那样稳定，但是如果碱基互补的整个区域很长（>100bp，图 6.10），则可以容忍很大程度的错配。

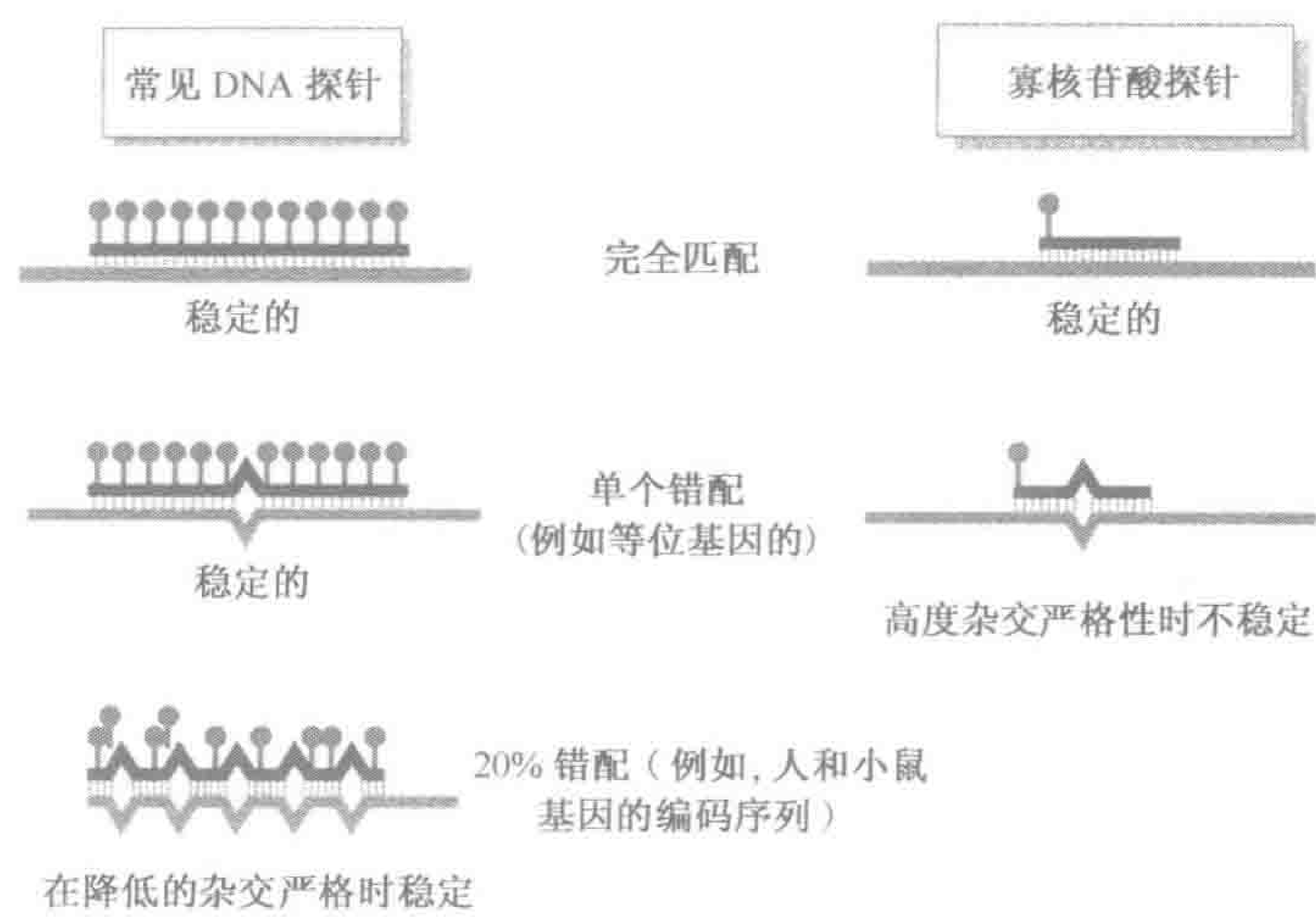


图 6.10　核酸杂交实验能够识别与一常规核酸探针差异相当大的，或者与一寡核苷酸完全相同的靶序列

增加 NaCl 的浓度和降低温度能够降低**杂交严格性**（hybridization stringency）而增加错配异源双链的稳定性。这意味着可以用一个特异的家族成员作为探针，通过杂交识别某一多基因家族或其他重复 DNA 家族相当变异的成员。另外，假设序列在进化过程中相当保守，也可以应用一个物种的这个基因序列为探针，识别其他相当变异的物种的同源序列。

也可以选择使杂交严格性最大化的杂交条件（如降低 NaCl 的浓度及提高温度），以促进错配异源双链的解离（变性）。如果碱基互补性区域很小，可以选择像使用寡核苷酸探针（一般为 15～20 个核苷酸）的杂交条件，即单个碱基错配便能使异源双链不稳定（节 6.3.1）。

6.2.2　DNA 浓度和时间（C_0t）的产物限定 DNA 重新联合动力学

当双链 DNA 变性（如通过加热变性），互补的单链可以重新联合形成双链 DNA 时，互补链重新联合的速度依赖于 DNA 的起始浓度。如果互补 DNA 序列浓度很高，那么任何一条单链 DNA 分子找到一条互补链并形成双链所花费的时间就会减少。**重新联合动力学**（reassociation kinetics）是用于测量互补单链分子能够彼此识别并形成双

链的速度的概念。有两个主要参数：特定 DNA 序列的起始浓度（C_o）（以每升核苷酸的摩尔数表示）及反应时间（t）（以秒表示）。由于再联合的速率与 C_o 和 t 成正比，因此 C_ot 值（通常粗略地称之为 C_ot 值）是一个有用的衡量标准。C_ot 值也依赖于重新联合的温度及一价阳离子的浓度而变化。结果，常用固定的参考值：一个 65℃的重新联合温度和一个 0.3M NaCl 的［Na^+］浓度。

为了促进探针-靶核酸的形成，大多数杂交实验使用超出探针过量的靶核苷酸。这是因为探针通常是同质的，常由一单个类型的克隆 DNA 分子或 RNA 分子组成，但靶核酸是典型地异质的，例如由基因组 DNA 或细胞总 RNA 组成。在后者，任何一个序列的浓度可能非常低，因此导致重新联合的速率也很低。例如，如果一个 Southern 印迹使用一个克隆的 β 珠蛋白基因作为探针来识别人类基因组 DNA 中的互补序列，那么后者将会以非常低的浓度存在（β 珠蛋白基因是单拷贝序列的一个例子，在此实例中仅代表人类基因组 DNA 的 0.00005%）。因此使用几微克靶 DNA 驱动该反应是非常必要的。相反，某些其他序列在基因组 DNA 中是高度重复的（第九章），这个非常高的 DNA 浓度一个相对快速的重新联合的时间。

因为探针结合的靶核酸量依赖于识别序列的拷贝数，所以杂交信号强度与识别序列的拷贝数成正比：单拷贝基因产生弱杂交信号，而高度重复 DNA 序列产生非常强的杂交信号。如果一个特定的探针是异质的，含有一低拷贝的兴趣序列（诸如一个特定的基因），与一个非常丰富的重复 DNA 序列混合，那么由低拷贝序列获得的弱杂交信号将会被来自重复 DNA 序列的强信号完全掩盖。可是，这种效应可通过**竞争杂交**（competition hybridization）克服（框 6.3）。

框 6.3　核酸杂交词汇表（各种方法见框 6.4）

等位特异性寡核苷酸杂交（allele-specific oligonucleotide hybridization）：应用短的、长约 20 个核苷酸，特异的寡核苷酸探针在严格的杂交条件下与单个等位基因杂交，一单个碱基的错配也能妨碍成功的杂交（图 6.11）。

复性（anneal）：如果两条单链核酸分享足够的碱基互补性，它们将形成一条双链 DNA。复性意味着允许两条单链之间形成氢键，与其相反的意思是**变性**（denaturation）。

碱基互补性（base complementarity）：两条单链核酸序列能够根据 Watson-Crick 碱基配对（节 1.2.1）形成一条 DNA 双链的程度。

竞争杂交（competition hybridization）（=**抑制杂交**，suppression hybridization）：含有某些重复 DNA 序列的探针经历一个预杂交步骤以阻断进入探针中重复序列通路的杂交反应。探针首先变性，在富含重复 DNA 的未标记 DNA 群体存在的情况下重新联合。结果，通过与互补重复 DNA 序列复性，探针内的重复元件被有效地清除，仅留下可用的非重复序列。

变性（denature）：通过破坏一双链 DNA 的各个单链间的氢键分离它们（通过加热它们或用一化学变性剂如甲酰胺处理它们）。

荧光原位杂交（FISH）：核酸通过附带在特定波长能够发出荧光的化学基团而被标记的任何**原位杂交**（*in situ* hybridization）反应。

异源双链（heterodoplex）：具有部分碱基互补性的两条单链序列复性形成的一条双链 DNA。异源双链能够以不同的方式出现。

框 6.3　核酸杂交词汇表（各种方法见框 6.4）（续）

等位基因异源双链（allelic heteroduplex）：变性然后冷却一单个二倍体样品或者来源于不同个体的 DNA 样品的混合物将使 DNA 序列稍有差异的两个等位基因的互补链形成几乎完全匹配的双链。

种内异源双链（paralogous heteroduplex）：变性然后冷却来源于一复杂真核细胞的一单个 DNA 样品，在非等位基因相关的 DNA 序列，诸如一个基因家族中或一非编码重复 DNA 家族紧密相关的不同成员，发生重新联合。

种间异源双链（interspecific heteroduplexe）：来源于理论上具有紧密相关基因的物种，例如人和小鼠的变性 DNA 样品被混合后单链允许重新复性。

同源双链（homoduplex）：两条具有完全碱基互补性的单链序列允许复性时形成的一条双链 DNA。

杂交实验（hybridization assay）：应用一已知核酸（**探针**，probe）寻找一异源 DNA 序列集合（**靶**，target）中相关序列的一种检测反应。在一标准的杂交实验中，探针被标记，通常位于溶液中，而靶序列未被标记，通常结合于一固体支持物，反向杂交与之相反。

原位杂交（*in situ* hybridization）：探针与来源于固定在载玻片上的细胞或染色体的核酸杂交的一个杂交反应。例如来源于中期染色体的变性 DNA（染色体原位杂交，节 2.4.2）或切片组织制备的 RNA（组织原位杂交）（节 6.3.4）。

熔解温度（melting temperature）（T_m）：一个核酸双链稳定性的衡量标准，它是与观察到的双链形式转换成单链形式的中点温度相一致的温度。方便的是，这种转换可通过测量在 260nm 波长处 DNA 的光密度完成（图 6.9）。

探针（probe）：在杂交实验中，用于查询一个复杂的异源核酸群的一个已知的核苷酸群。起始可以是双链，但工作时作为探针必须变性为单链。探针通常是标记的，并位于溶液中，但参见反向杂交实验。

重新联合（reassociate）：经过一个先前的变性步骤后重新复性。

重新联合动力学（reassociation kinetics）：互补 DNA 链重新联合的速率，高度重复 DNA 链重新联合快，而简单拷贝序列重新联合的要慢些。

反向杂交实验（reverse hybridization assay）：一个靶序列被标记，常位于溶液中，而探针未标记，常结合于一固相支持物的实验。如包括反向斑点杂交实验和微阵列杂交。

消减杂交（subtraction hybridization）：一种以杂交为基础的、识别存在于一个群体（正群体）而不存在于另一个密切相关群体（负群体）的已知核酸序列的方法。杂交设计为具有极其过量的（－）群体并因此作为**驱动 DNA**（driver DNA）起作用，确定（＋）群体中所有相关序列作为异源双链被清除，留下仅发现于（＋）群体的期望序列。

靶（target）：杂交实验中用一个已知的，通常是同源的核酸来查询的一复杂异源核酸群。通常是未标记，并结合于一个固相支持物，但在反向杂交实验中，靶核酸是标记的且位于溶液中。

6.2.3　可应用的多种多样的核酸杂交实验

早期的核酸杂交实验利用溶液杂交，涉及探针和靶核酸水溶液混合。然而在复杂基因组中单拷贝序列非常低的浓度意味着重新联合的时间必然很慢。一种广泛使用的提高重新联合速度的方法是通过人为地提取水分子以提高水溶液中总的 DNA 浓度（如加入高浓度的聚二乙醇）。

一种可选择的易于重新联合分子检测的溶液杂交涉及将靶 DNA 固定于一个固体支持物上，如硝酸纤维素膜或尼龙膜，单链 DNA 很容易与这两者结合。然后，标记探针结合于固定的靶 DNA，紧接着除去含有未结合探针 DNA 的溶液，广泛洗膜，晾干以备检测。

这是目前使用的标准核酸杂交实验的基础。然而，最近**反向杂交实验**（reverse hybridization assay）也变得非常普遍。在这些实例中，探针群是未标记的，并固定于固体支持物，而靶核酸被标记且存在于水溶液中。因此需要指出，探针和靶序列之间的区别不是主要依据哪是标记群，哪是未标记群。反而，重要的需要考虑的事是靶 DNA 应该是复杂的、不完全了解的、探针（分子特性已知）试图去查询的群体。依据探针和靶序列的性质和形式，可以设计多种多样的核酸杂交实验（框 6.4）。

框 6.4 标准和反向核酸杂交实验

标准实验	溶液中标记的探针	结合于固体支持物的未标记的靶序列
斑点印迹（图 6.11）	任何标记的 DNA 或 RNA，但通常是寡核苷酸	复杂的 DNA 或 RNA 群；未按大小分离，而直接点于膜上
Southern 印迹（图 6.12 示方法；图 7.9，18.12B，18.12C 举例）	任何	通常是复杂的基因组 DNA（但也可能是单独的 DNA 克隆）；限制性核酸酶消化，按大小分离，然后转移至膜上
Northern 印迹（图 6.13）	任何	复杂 RNA 群（如，细胞总 RNA 或 $polyA^+$ RNA）已按大小分离，后转移至膜上
染色体原位杂交（例如，图 2.16－2.18，14.12）	通常是一个标记的基因组克隆	显微镜载玻片上裂解细胞染色体（常为中期）的 DNA
组织原位杂交（图 6.15）	通常是一个标记的反义核糖探针或寡核苷酸	显微镜载玻片上固定的组织切片细胞内 RNA
集落印迹	任何	琼脂上铺板后分离，再转移至膜上的细胞集落
斑-隆起	任何	琼脂上铺板后分离并转移至膜上的噬菌体感染细菌集落
网格克隆杂交实验（图 6.17）	任何	自动点样于位于几何图形陈列的膜上的克隆
反向实验	**溶液中标记的靶序列**	**结合于固体支持物的未标记探针**
反向斑点印迹	复杂 DNA	点样于膜上的寡核苷酸
DNA 微阵列或预先合成的寡核苷酸微阵列（图 6.19）	复杂 DNA	自动点样于显微镜载玻片的 DNA 克隆或寡核苷酸
寡核苷酸微阵列（图 17.18）	复杂 DNA	在玻片上合成的寡核苷酸

6.3 使用克隆的 DNA 探针筛查未克隆的核酸群进行核酸杂交实验

分子遗传学的许多应用涉及获得一单个的 DNA 克隆并将它作为一个杂交探针，在

一个复杂的未克隆的 DNA 或 RNA 靶序列中筛查相关序列的存在。有时实验仅限于简单地检测与探针相关的序列存在与否。在其他的例子中能够获得关于互补序列的大小，它们的亚染色体定位或它们在特异组织或细胞群中定位的有用信息。

6.3.1　斑点印迹杂交——常使用等位基因特异性寡核苷酸探针的一种快速筛查方法

斑点印迹（dot-blotting）的一般程序包括获取靶 DNA（如人类全基因组 DNA）的水溶液，并简单地点样于一硝酸纤维素膜或尼龙膜上，然后使其干燥。改变的技术——**狭缝印迹**（slot-blotting）包括从一适宜的模板中吸取 DNA 使其通过一单独的狭缝。在这两种方法中，靶 DNA 序列的变性，要么通过先前热处理，要么将包含它们的滤膜置于碱溶液中。现在，固定于膜上的变性靶 DNA 序列置于含有单链标记探针序列的溶液中（标记通常为 ^{32}P 以便于检测）。给予足够的时间用于探针－靶序列异源双链形成后，轻轻倒出探针溶液，洗膜以清除可能与滤膜非特异性结合的过量探针。然后晾干，接触放射自显影胶片。

斑点印迹一个有益的应用，包括对等位基因间甚至只有一单个核苷酸替代差异的区分。为实现此目的，需要从跨越变异核苷酸位点的序列构建**等位基因特异性寡核苷酸**（allele-specific oligonucleotide，ASO）探针。典型的 ASO 探针为 15～20 核苷酸长，正常使用于探针和靶序列间形成的 DNA 双链仅在它们之间具有完全碱基互补性才稳定的杂交条件：探针和靶序列间一单个碱基错配足以使短的异源双链不稳定（图 6.10）。典型地，这包括设计寡核苷酸使得等位基因间单个核苷酸的差异存在于寡核苷酸序列的中心部分，从而使一个错配双链的热力学不稳定性最大化。这种区别可用于各种研究及诊断目的。尽管 ASO 能用于常规的 Southern 印迹杂交（见下文），但将其用于斑点印迹实验更方便（图 6.11）。

另一种 ASO 斑点印迹方法使用一**反向斑点印迹**（reverse dot blotting）方法。这意味着寡核苷酸探针未被标记且固定于一滤膜或膜上，而靶 DNA 在溶液中被标记并提供。标记靶 DNA 与膜上某一特定寡核苷酸的阳性结合意味着靶 DNA 具有那个特异序列。这种方法及相关的 DNA 微阵列方法具有许多诊断性应用。

6.3.2　Southern 和 Northern 印迹杂交检测通过凝胶电泳按大小片段分离的核酸

Southern 印迹杂交

在此方法中，靶 DNA 使用 1 个或多个识别序列常为 4～6bp 长的限制性内切核酸酶消化，产生几百或数千 bp 长的片段。限制片段经琼脂糖凝胶电泳按大小片段分离，变性并转移至一硝酸纤维素膜或尼龙膜上用于杂交（图 6.12）。在电泳过程中，由于磷酸基团而带负电荷的 DNA 片段与负极相排斥而向正极迁移，并滤过孔性凝胶，片段越小泳动越快。对于 0.1～30kb 长的片段来说，迁移率取决于片段长度而几乎根本不依赖于碱基组成。因此，在此大小范围内的片段在一常规琼脂糖凝胶电泳系统中是按其大小进行分离的。

Southern 印迹杂交在哺乳动物遗传学中的一个重要应用是使用一个探针去识别可能属于同一基因组的相关序列（一个进化相关基因或 DNA 序列家族的其他成员）或

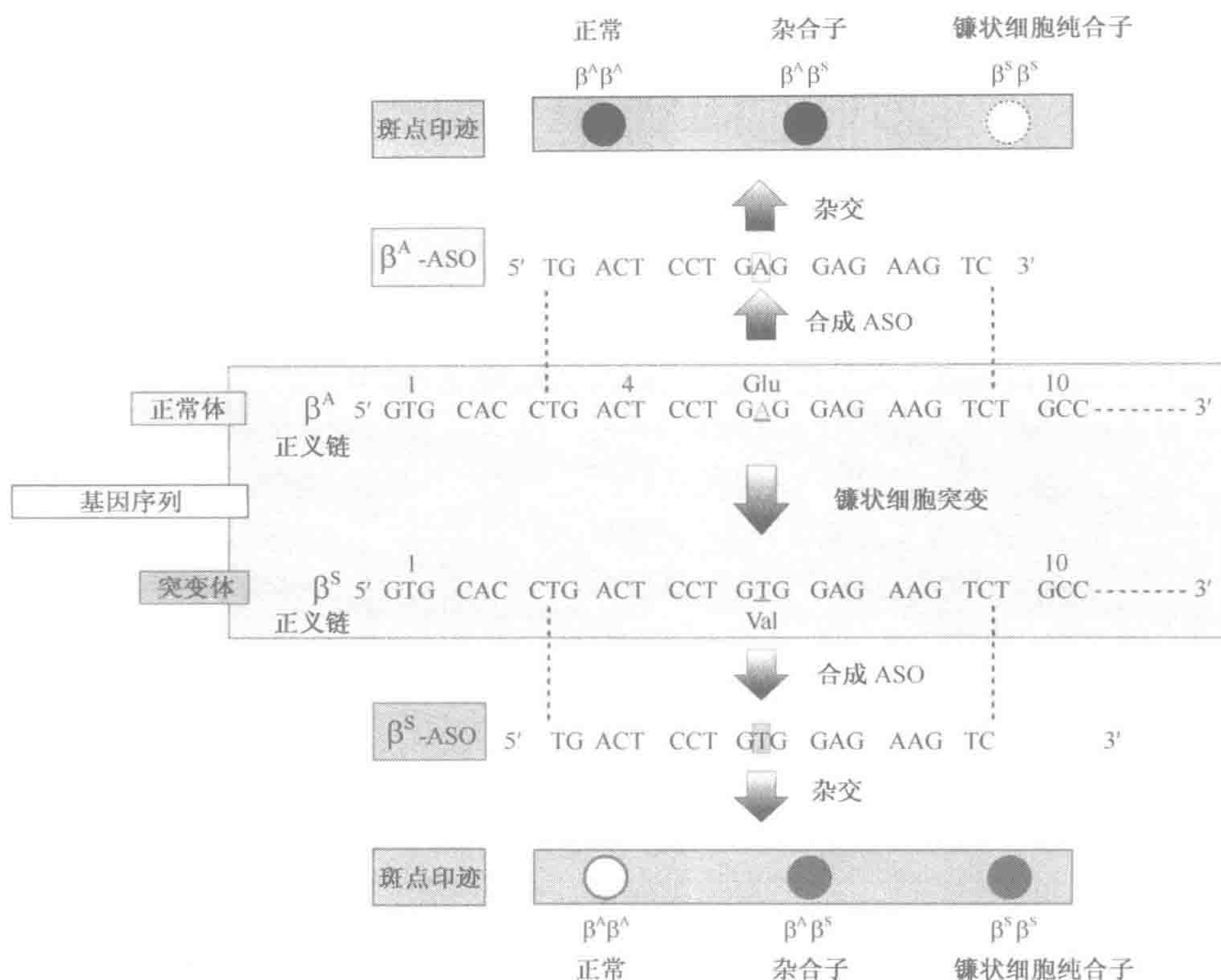


图 6.11　等位基因特异性寡核苷酸（ASO）斑点印迹杂交可以识别镰状细胞突变的个体

顶部图解的斑点印迹杂交显示正常 β 珠蛋白的等位基因特异的 ASO 探针杂交结果（β^A ASO；显示于紧邻的下方）。正常个体及杂合子的结果为阳性（实心圆圈），而镰状细胞纯合子的结果为阴性（虚线，未填充圆圈）。底部斑点印迹显示镰状细胞 β 珠蛋白等位基因特异的 ASO 探针杂交结果（β^S ASO；显示于紧邻的上方）。在本例中，镰状细胞纯合子和杂合子的结果为阳性，而正常个体的结果为阴性。本例中 β^A ASO 和 β^S ASO 设计为 19 个核苷酸长，围绕镰状细胞突变位点分别选自正义链 β^A 和 β^S 珠蛋白基因序列的第 3～第 9 密码子。后者在 β 珠蛋白基因第 6 密码子处有一单个核苷酸替代（A→T），导致一个 GAG（Glu）→GTG（Val）替代（见中间序列）。

其他基因组的相关序列（如种间同源基因，即用作探针的基因的一个直接等价物），一旦一个新分离的探针显示出与其他未定性的序列相关，那就可以尝试通过筛查基因组 DNA 文库（节 5.3.4）来分离这一家族的其他成员。另外，也可对不同物种的基因组 DNA 样品进行筛查（**动物印迹**，zooblot）以识别不同物种间的保守序列（图 7.9）。因为编码序列是相对高度保守的，所以这是识别编码 DNA 的一条途径。

Northern 印迹杂交

Northern 印迹杂交是 Southern 印迹的一个改变形式，它所用的靶核酸是未消化的 RNA 而不是 DNA。这种方法的主要应用是获取特定基因表达模式的信息。一旦克隆了一个基因，就能够以它作为探针，与不同泳道内从多种不同组织分离的 RNA 样品进行 Northern 印迹杂交（图 6.13）。获得的数据能够提供基因表达的细胞类型范围以及转录物相对的丰度。另外，通过揭示不同大小的转录物，可为不同的异构体提供证据（例如

图 6.12　Southern 印迹杂交检测通过凝胶电泳按大小分离的靶 DNA 片段

源于选择性启动子、剪接位点、多聚腺苷酸化位点等)。

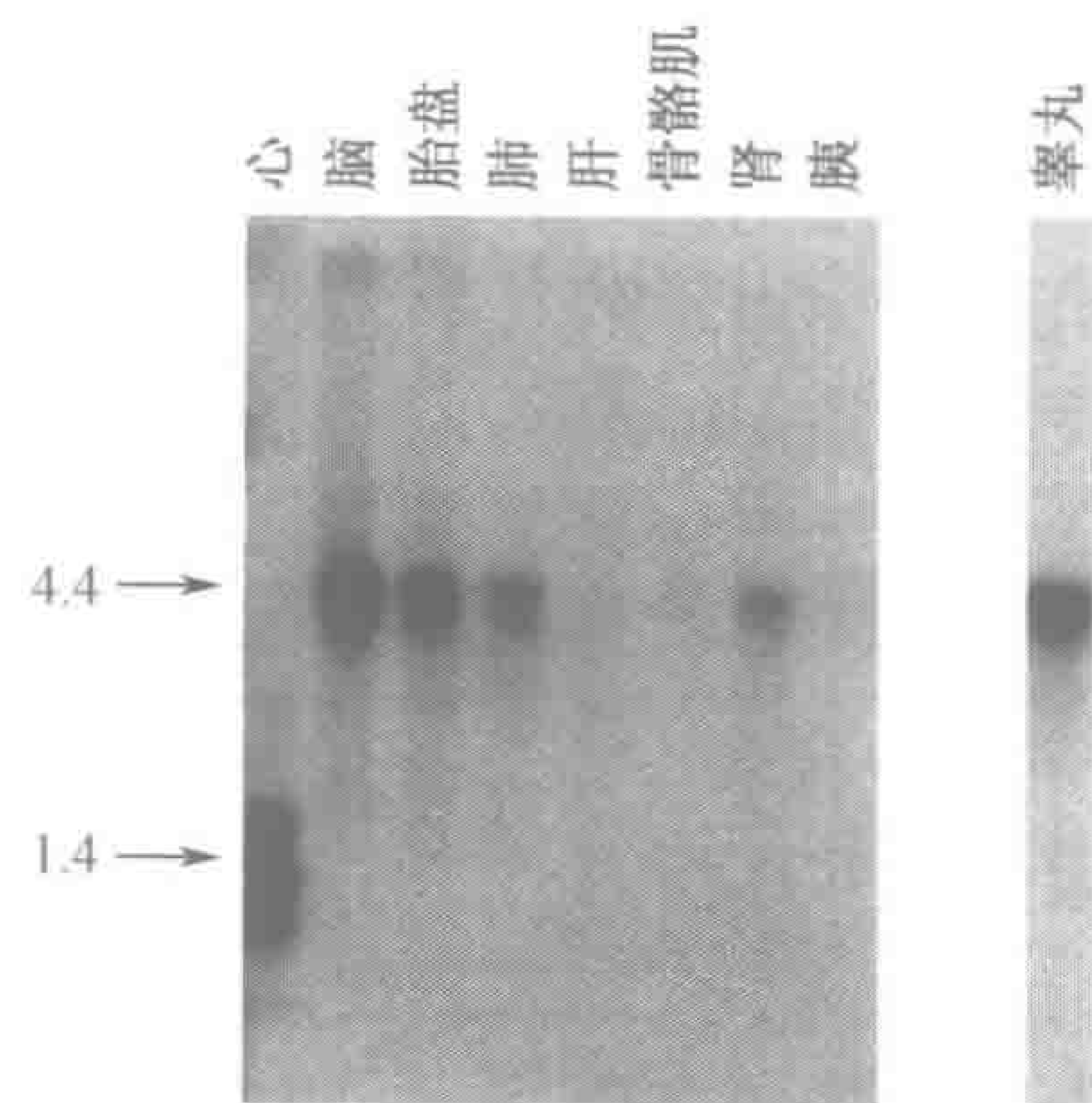


图 6.13　Northern 印迹杂交用于评价一个基因总的表达模式

Northern 印迹包括按照大小分离总 RNA 样品［或纯化的 poly（A⁺）mRNA］，转移至膜上并与一个适宜的标记核酸探针杂交。本例显示了来源于 *FMR1*（脆性 X 智力低下综合征）基因的一个标记 cDNA 探针的使用。在脑和睾丸（4.4kb）检测到最高表达水平，在胎盘、肺和肾分别为降低的表达，在心脏中出现多个较小的转录物。经 Nature Publishing Group 同意，再引自 Hinds 等（1993）。

6.3.3 脉冲电场凝胶电泳扩展 Southern 印迹杂交至包括很大的 DNA 分子的检测

由于滤网效应，标准的琼脂糖凝胶电泳能够分辨有限大小范围的 DNA 片段，从 100bp 到大约 30kb：DNA 分子通过琼脂糖凝胶的孔隙，而小分子能够更快地迁移通过孔隙。然而，对于超过某一特定大小的 DNA 片段，滤网效应不再有效，超过 40kb 的 DNA 片段的分辨率将极度受限。因为许多哺乳动物基因和其他功能序列单位很大，所以需要一种可选择的电泳方法来分离很大的 DNA 片段。

脉冲电场凝胶电泳（pulsed field gel electrophoresis，PFGE）是近来一种改进的琼脂糖凝胶电泳，它能分辨大约 20kb 至几个 Mb 长度范围的 DNA 片段。哺乳动物染色体中包含的很大的 DNA 分子——典型地数百 Mb 长——不能应用这种方法按大小分离，但是专用的限制性核酸酶可以切割脊椎动物 DNA 相当罕见地产生的大限制片段，这些片段可以通过 PFGE 按大小分离。这些酶，有时称为**“罕见切割限制性内切核酸酶”**（rare-cutter restriction endonuclease），通常识别含有一个或更多 CpG 二核苷酸的富含 GC 的识别序列。因为 CpG 二核苷酸在脊椎动物 DNA 以低频率出现，所以人类 DNA 和其他脊椎动物 DNA 具有比较少的被这种切割含有 CpG 序列的限制酶所识别的序列（表 6.3）。

表 6.3　“罕见切割”限制性内切核酸酶举例

酶	来源	切割序列：CG=CpG；N=A,C,G 或 T	人类 DNA[a] 中平均期望的片段大小(kb)
Sma Ⅰ	*Serratia marcescens*	CC**CG**GG	78
BssH Ⅱ	*Bacillus stearothermophilus*	G**CGCG**C	390
Sac Ⅱ	*Streptomyces lividans*	C**CGCG**G	390
Sfi Ⅰ	*Streptomyces fimbriatus*	GGCCNNNNNGGCC	400
Not Ⅰ	*Norcadia otitidis-caviarum*	G**CG**GC**CG**C	9766

a 假设 40%(C+G)，期望 CpG 频率为 20%。

正常制备的基因组 DNA 不适于 PFGE，因为裂解细胞和纯化 DNA 的过程导致产生引起相当多 DNA 断裂的切力。相反，以这种方式分离 DNA 可使大分子人为地断裂为最小化，然后用适宜的罕见切割限制性内切核酸酶消化。为了制备高分子量 DNA，细胞样品，例如白细胞，将其与熔化的琼脂糖混合，然后移至一凝块模具孔中，并使其冷却，结果细胞包埋于固体琼脂糖块内（图 6.14）。移出琼脂糖凝块，与水解酶共同孵育，水解酶通过琼脂糖的小孔弥散并消化细胞成分，但留下真正完整的高分子量染色体 DNA。然后含有纯化的高分子量 DNA 的单个凝块能够与含有罕见切割限制性内切核酸酶的缓冲液共同孵育。

为了按大小分离包埋于琼脂糖凝块内的大限制片段，将凝块放置于 PFGE 装置所含有的一块琼脂糖凝胶一端的孔内。与传统的凝胶电泳一样，带负电荷的 DNA 与负极相排斥并在电场中迁移。然而，在 PFGE 运行过程中，凝胶和电场的相对方向是周期性转变的，典型地是通过安装一个开关来传送短暂的脉冲，交替激发两个不同方向的脉

冲场（图 6.14）。这种技术的改变形式使用一单个电场，但具有周期性极性转换（**电场倒转凝胶电泳**，field inversion gel electrophoresis）或周期性凝胶或电极的旋转。

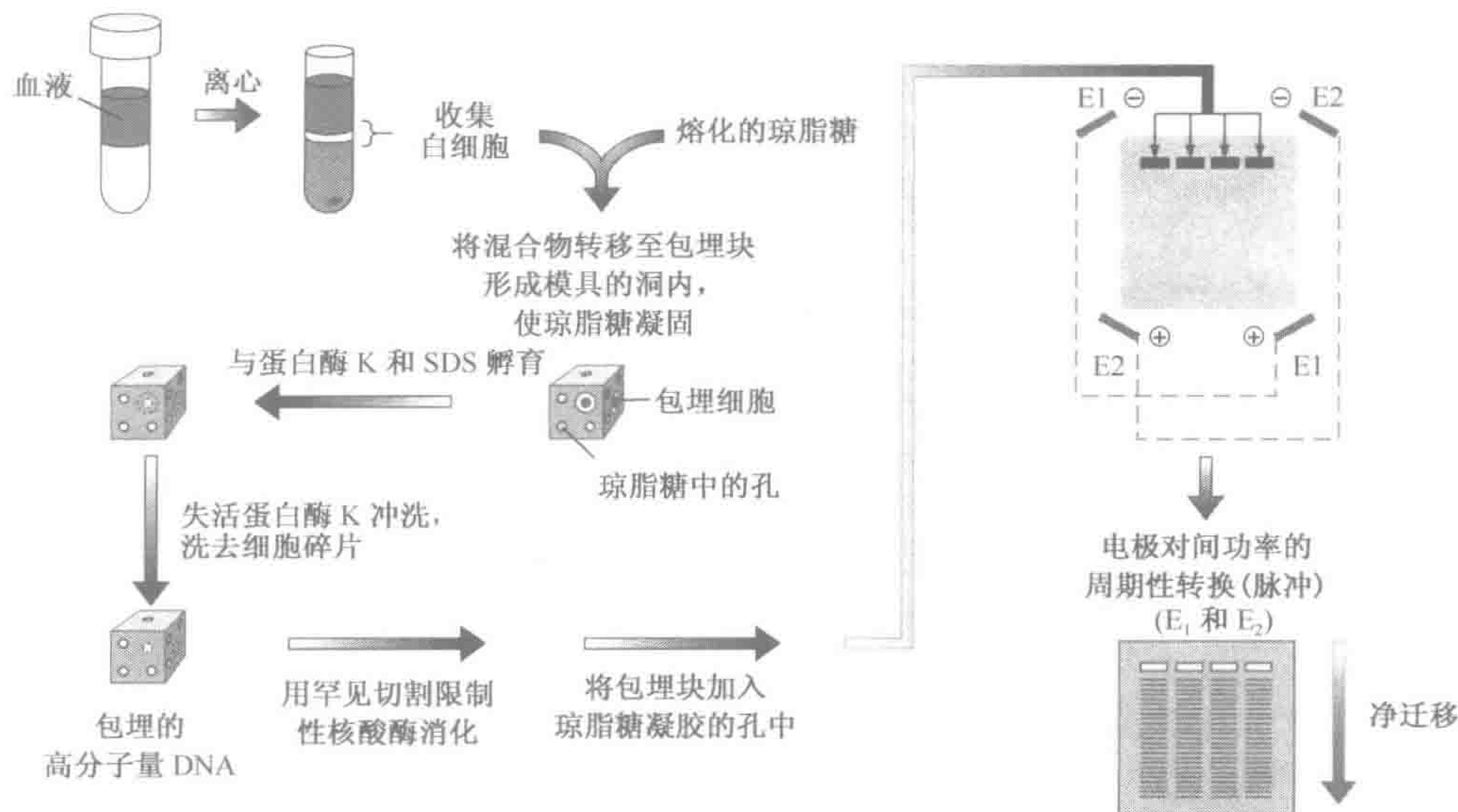


图 6.14　通过脉冲电场凝胶电泳从血细胞中分离高分子量 DNA 片段

每种改变方法的共同点是一个不连续电场的原理，使得 DNA 分子在通过凝胶过程中间歇地被迫改变它们的构象与迁移方向。一个 DNA 分子改变其构象及使它自己重新适应新电场的方向所花费的时间是严格地依赖于其大小的。结果最大几个 Mb 大小的 DNA 片段——包含来自整个酵母染色体的完整 DNA——能被有效地分离（Schwartz and Cantor，1984）。

6.3.4　在原位杂交中探针与一染色体标本的变性 DNA 或固定于一载玻片上组织切片 RNA 进行杂交

染色体原位杂交

为定位基因和其他 DNA 序列的一个简单方法是利用适宜一个标记的 DNA 探针与已经原位变性的染色体 DNA 杂交。为了达到此目的，需制备空气干燥的显微镜玻片的染色体标本（通常为来源于外周血淋巴细胞或成淋巴细胞样细胞系的中期或前中期染色体）。用核酶（RNase）及蛋白酶 K 处理产生部分纯化的染色体 DNA，通过置于甲酰胺中变性。然后加入含有一标记核酸探针的溶液，覆盖盖玻片，变性 DNA 可用于原位杂交。根据应用的特定技术，染色体上的显带可安排在杂交步骤之前或之后。结果，为了鉴定探针所识别的 DNA 序列在图谱上的位置，必须在除去多余探针后才获及与染色体带型相关的信号。由于**荧光原位杂交**（fluorescence *in situ* hybridization）技术的使用（节 2.4.2），染色体原位杂交已彻底改革。

组织原位杂交

在此方法中，一个标记的探针与组织切片上的 RNA 杂交（Wilkinson，1998）。使用一低温恒温器，组织切片可由石蜡包埋或冰冻切片制备，然后固定于玻璃载玻片上。包含探针的杂交混合物置于载玻片上的切片，覆以盖玻片。典型地杂交混合物含有浓度为 50%的甲酰胺，以降低杂交温度及使蒸发问题最小化。

尽管双链 cDNA 可用于探针，但更推荐应用单链互补 DNA 探针（**核糖探针，**riboprobe）。起始单链探针的敏感性普遍高于双链探针；可能是由于一部分变性的双链探针再复性，形成探针同源双链。与一个基因的 mRNA 互补的 cRNA 核酸探针称为**反义核糖探针**（antisense riboprobe），可通过在一个适宜的载体如 pSP64 中反向克隆一个基因而获得（图 6.3）。在这些实例中，噬菌体聚合酶将从与体外正常转录的链相反的 DNA 链合成被标记的转录物。对这一反应有用的对照包括有义核糖探针，除了一个基因的两条 DNA 链均被转录这种罕见的情况，有义核糖探针不与 mRNA 杂交。

探针的标记通过选择性同位素，如著名的 ^{35}S 或非同位素来进行。在前一种情况，应用放射性自显影方法杂交探针是可见的，通常仅使用**暗视野显微镜**（dark-field microscopy）可见银粒子的定位（直射光不能到达物镜，相反，光的照明线从侧面被引导，使得仅有散射光进入显微镜透镜，在黑色背景下信号以一个明亮的目标出现）。然而**亮视野显微镜**（bright-field microscopy）（通过光的直接透射通过样品获得图像）提供更好的信号检测（图 6.15）。荧光标记是一个受欢迎的非同位素标记方法，通过荧光显微镜完成检测（框 6.2，图 6.5）。

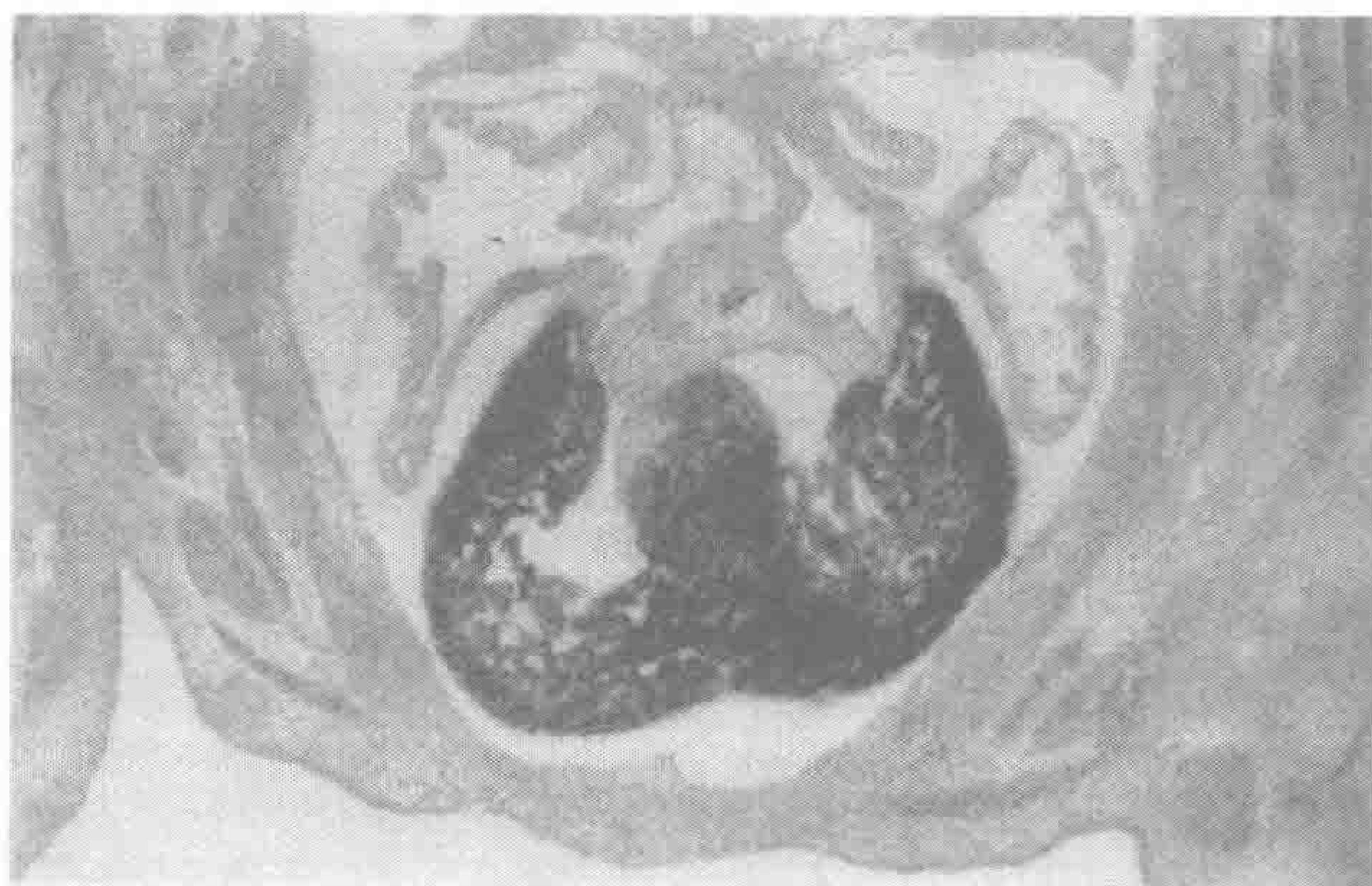


图 6.15　组织原位杂交提供高分辨基因表达模式

本例显示应用 ^{35}S 标记 β 肌球蛋白重链反义核糖探针，与 13 天胚胎小鼠组织横切切片杂交模式。暗区代表强标记，特别在心脏的心室明显。由英国 University of Newcastle uponTyne David Wilson 博士提供。

6.4　使用克隆靶 DNA 及微阵列的杂交实验

前面章节描述的一些技术（例如 Southern 印迹杂交和斑点印迹杂交）也可用于研究克隆的 DNA 和未克隆的 DNA。然而，下面两节描述的技术用于分析克隆的 DNA。

另外，新近发展的、非常强有力的微阵列技术在第三节描述。

6.4.1　集落印迹及斑一隆起杂交是筛查分离的细菌集落或菌斑的技术

正如节 5.3.5 所述，含有重组 DNA 的细菌或其他适宜宿主细胞的集落一般能通过插入物失活一个标记载体基因（如β半乳糖苷酶或抗生素耐药基因）而被选择或识别。然而，如果期望的重组 DNA 含有与一个可用的核酸探针密切相关的 DNA 序列，那么就可以通过杂交进行特异性检测。对于用来增殖质粒重组体的细菌细胞来说，细菌集落允许生长于琼脂表面，然后通过表面接触转移至硝酸纤维素膜或尼龙膜上，这一过程称为**集落印迹**（colony blotting）（图 6.16）。另一种方法是细胞混合物被铺展于一营养琼脂表面的硝酸纤维素膜或尼龙膜上，集落在膜的表面直接形成。在任一方法中，在与一标记的核酸探针杂交前，膜随后置于碱液中使 DNA 变性。

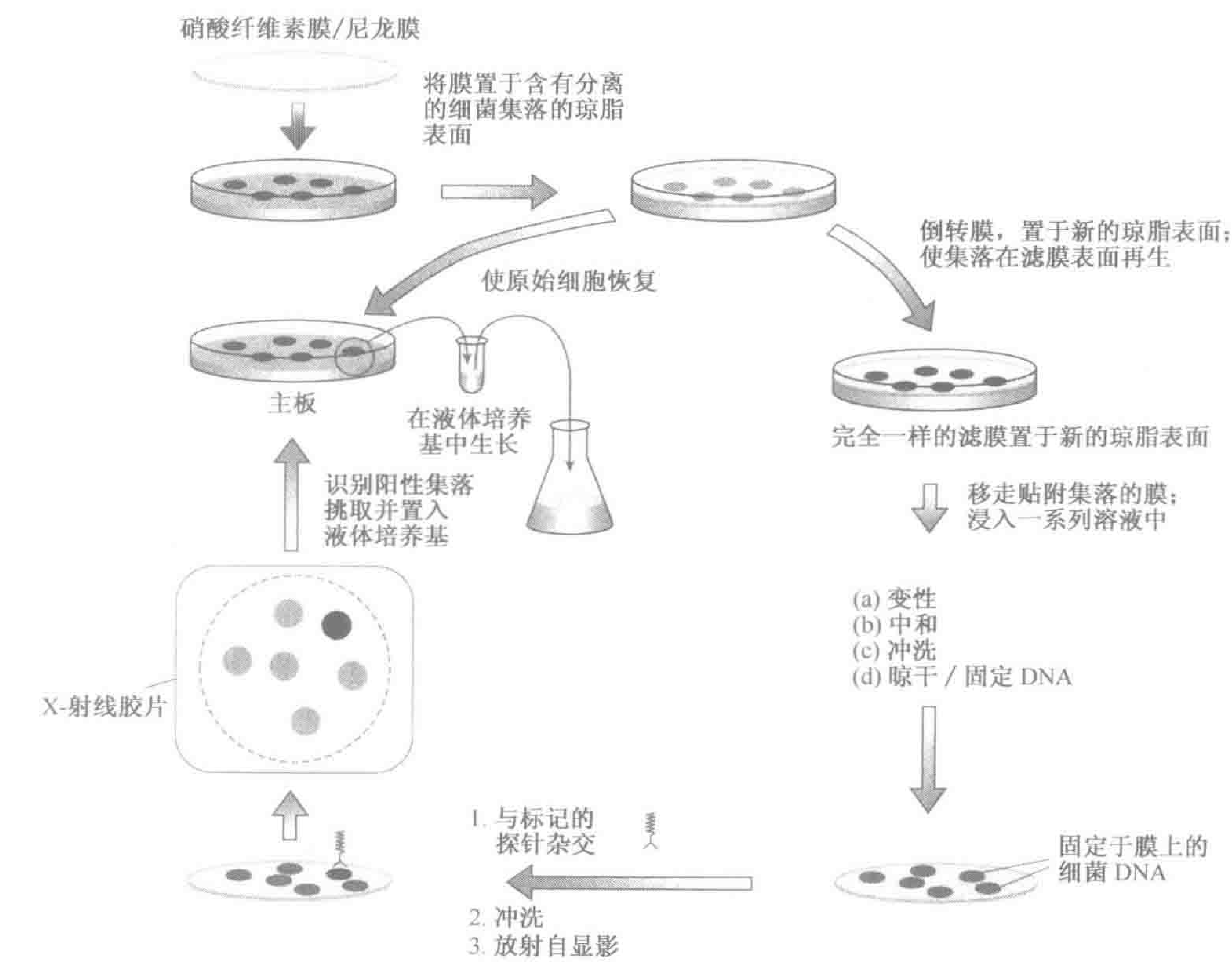


图 6.16　集落印迹杂交指在与一标记核酸探针杂交前将集落复制在一耐用的膜上

这种方法广泛用于鉴定含有 DNA 重组的集落，应有一个可用的适宜标记的探针。

杂交后清除探针溶液，广泛地冲洗滤膜、晾干、并应用 X 射线胶片使其放射自显影。为了鉴定含有与探针相关的 DNA 的集落，强放射信号的位置反过来与含有集落原始格局的主板相关联。随后在 DNA 提取及纯化 DNA 重组之前，这些集落能被单个挑取，并在培养基内扩增。

当使用噬菌体载体时，相似的过程也是可能的，随着噬菌体裂解细菌细胞，将形成

含有残留的噬菌体颗粒的斑。将硝酸纤维素膜或尼龙膜按上述方法置于琼脂板表面。当它们从板上移去时，将会形成一个菌斑中噬菌体物质的真实的拷贝，即所谓的斑—隆起。随后滤膜的加工与图 6.16 的图解一致。

6.4.2 转化细胞克隆或 DNA 克隆的网格高密度阵列极大提高了 DNA 文库筛查的效率

一旦构建复杂的 DNA 文库成为可能，就需要更为有效的克隆筛查的方法。与其在标准的集落印迹中简单地将细胞集落铺于一个细胞培养皿，并将它们转移至膜上，不如选取单个集落，并将它们转移至具有**高密度网格阵列**（high-density gridding array）格式大的膜上。

通过应用**自动网格设备**（robotic gridding device）使阵列的产生过程极大地简单化，这种设备能够通过吸取位于微滴定盘内的克隆点样于膜上预先确定的线性坐标而进行需要的自动点样。结果产生的高密度克隆滤膜允许快速有效地进行文库筛查（图 6.17)，并且能被拷贝和分配给全球内的许多实验室。

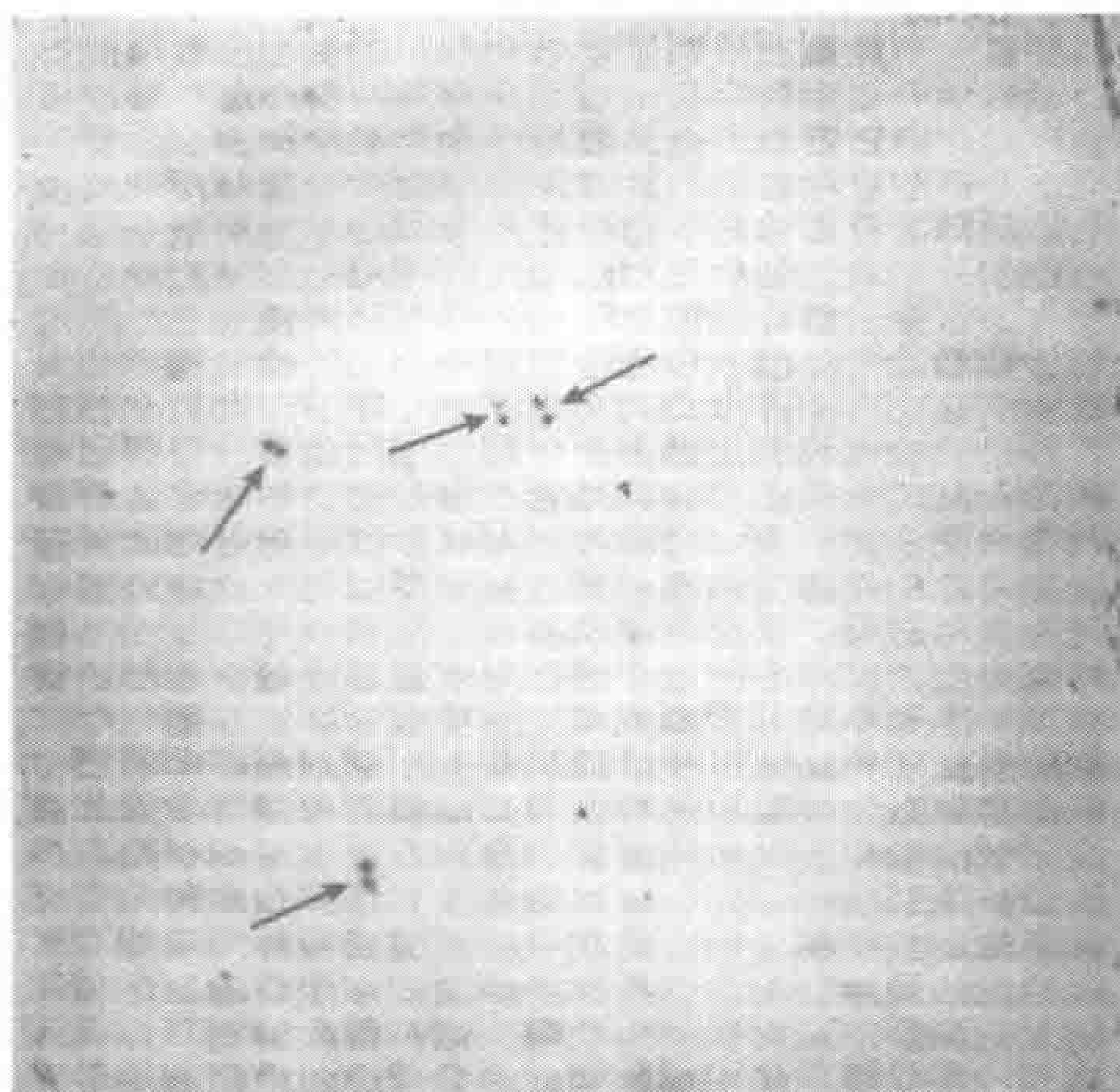


图 6.17　网格克隆杂交滤膜促进人类基因组的物理绘图

该图显示了一张包含人类 YAC 克隆的膜的放射自显影图片（即含有人类 YAC 的各个酵母克隆总 DNA)。膜上含有总计 17664 个克隆，被分隔在单位网格为 6×6 克隆的阵列中。杂交信号包括使用总酵母 DNA 的 ^{35}S 标记探针由所有克隆得到的弱信号以及使用 ^{32}P 标记的独一无二的性染色体探针（DX-YS646）得到的强杂交信号。原始图片来源 Cambrige 的 Sanger 中心 Mark Ross 博士。再引自 Ross and Stanton (1995) in: *Current Protocols in Human Genetics*, Vol. 1，经 Wiley-Liss，Inc.，John Wiley&Sons 的附属机构允许使用。

6.4.3 DNA 微阵列技术极大地扩展了核酸杂交的能力

最近发展的 **DNA 微阵列**（DNA microarray）技术由于其巨大的微型化和自动化能力，已扩展了杂交实验技术（Schena *et al.*,1998)。虽然使人联想起以滤膜为基础的阵列，但微阵列的构建仍包括完全不同的步骤。它的表面是典型的化学处理的载玻片，而不是多孔的膜，尽管为了尝试结合更多的 DNA 及增加敏感性，最近倾向于采用更加多孔的底物，如硝酸纤维素膜—包被的玻璃表面。根据核酸样品如何产生及如何传递至微阵列的不同，可分为两种完全不同类型的微阵列技术：

▶ **预先合成核酸的微阵列**。在这里，存在于微阵列的核酸是预先合成的（通常由不同 DNA 克隆的集合组成，但原则上它们可以是预先合成的集合）。在这种情况中微阵列的构建意味着单个 DNA 克隆或寡核苷酸被点样于一显微镜载玻片表面的各个位置，此位置在微型网格通过精确的 X、Y 坐标指定（图 6.18A）。传递样品的高精确化需要相当高级的自动接触印刷设备，而制造这类设备的详细说明已公布于：http://cmgm.stanford.edu/pbrown/mguide/；

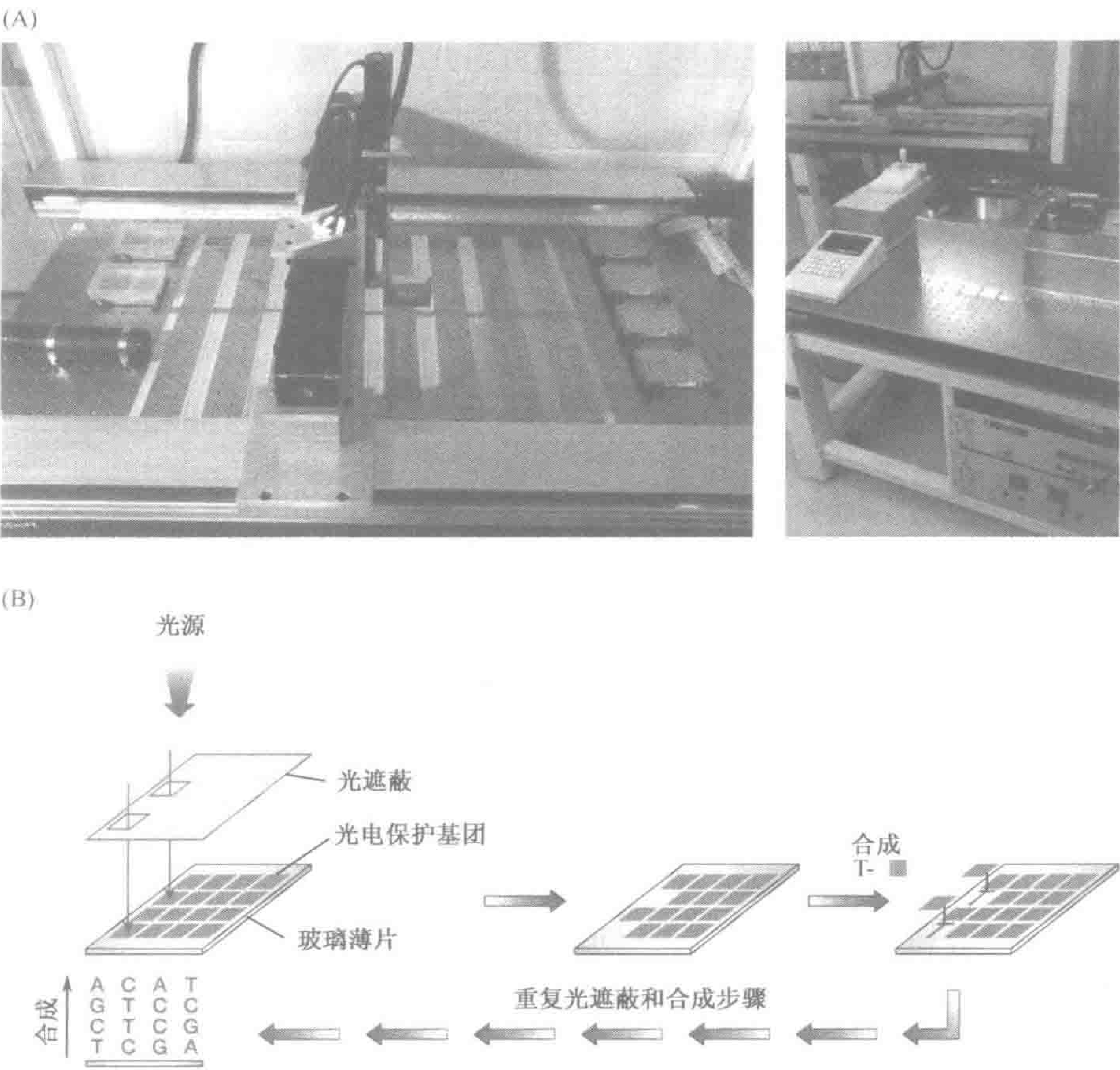


图 6.18　DNA 和寡核苷酸微阵列的构建

（A）**自动点样构建的 DNA 微阵列**。左：一个微阵列自动机：具有一个工作台的装置，含有 160 个载玻片，4 个微滴定盘，两个冲洗站和干燥器。右：一个显示光学工作台激光扫描仪，为激光和光电倍增管冷却装置、Ludi stage 及透镜提供能量供应（详见 Cheung *et al.*，1999）。机器对样品的微量点样能够通过点样针与固体表面（显微镜载玻片）间的物理接触，或者通过标准印刷中所用的喷墨方法完成（样品装入具有压电附件的微型喷嘴内，利用电流将精确量的液体从喷嘴内排至基质上），图片由 Albert Einstein College of Medicine 的 Aldo Massimi，Kaju Kucherlapati 和 Geoffrey Childs 友好提供。经 Nature Publishing Group 允许，再印至 Cheung 等（1999）. Nature Genet. 21（Suppl.），15～19。（B）**寡核苷酸微阵列的构建**结合了光刻及寡核苷酸原位合成。寡核苷酸通过连续的步骤原位合成，起始于锚定在玻璃薄片表面的 3′单核苷酸。光刻必须修正具有光电保护基团的玻璃薄片（当这一基团暴露于光时能够被清除）以及仔细构建的光遮蔽的使用，使光通过到达仔细选择的空间坐标。对于薄片上通过光遮蔽接受光的那些区域，除去光电保护基团可进行一个新的合成步骤。在本例中，胸腺嘧啶表明与一保护性光电基团偶联。

▶ **原位合成的寡核苷酸微阵列**。这种方法是 Affymetrix 公司首创的，典型地涉及来自半导体工业的照相平板印刷术与寡核苷酸合成化学的结合。在这种情况下，数千种不同寡核苷酸通过每次增加一个核苷酸的系列连续合成步骤原位聚集于载玻片的表面。这一过程需要将单核苷酸共价偶联至一个连接分子——末端连有光电的保护基团（图 6.18B）。

光遮蔽用于确定哪个部位与光反应：在一个特殊位置开放光遮蔽将允许外部光源照亮，破坏光电保护基团。然后应用化学偶联反应在新的脱保护位置加入一特定类型的核苷酸，使用不同的遮蔽重复这一过程，而遮蔽隐藏的那些位置将受到保护不与光接触。通过这种方式可在预先确定的位置构建特异性核苷酸序列，而这一位置主要依赖于用于那个合成步骤的照相平板印刷术中所切割的孔的排列。与硅片生产的相似性造成 **DNA 芯片**（DNA chip）这一名称普遍用于此类微阵列，由于生产 DNA 芯片需要复杂的技术，所以需要从专门的公司购买 DNA 芯片。

就像在反向斑点印迹中一样（节 6.3.1），DNA 微阵列技术按照反向核酸杂交方法起作用。探针是固定于微阵列的未标记的一组核酸。尽管复杂，探针的量是已知，并用于查询靶序列，靶序列虽然由溶液中标记的核酸组成，但其来源于我们所需要的信息资源。

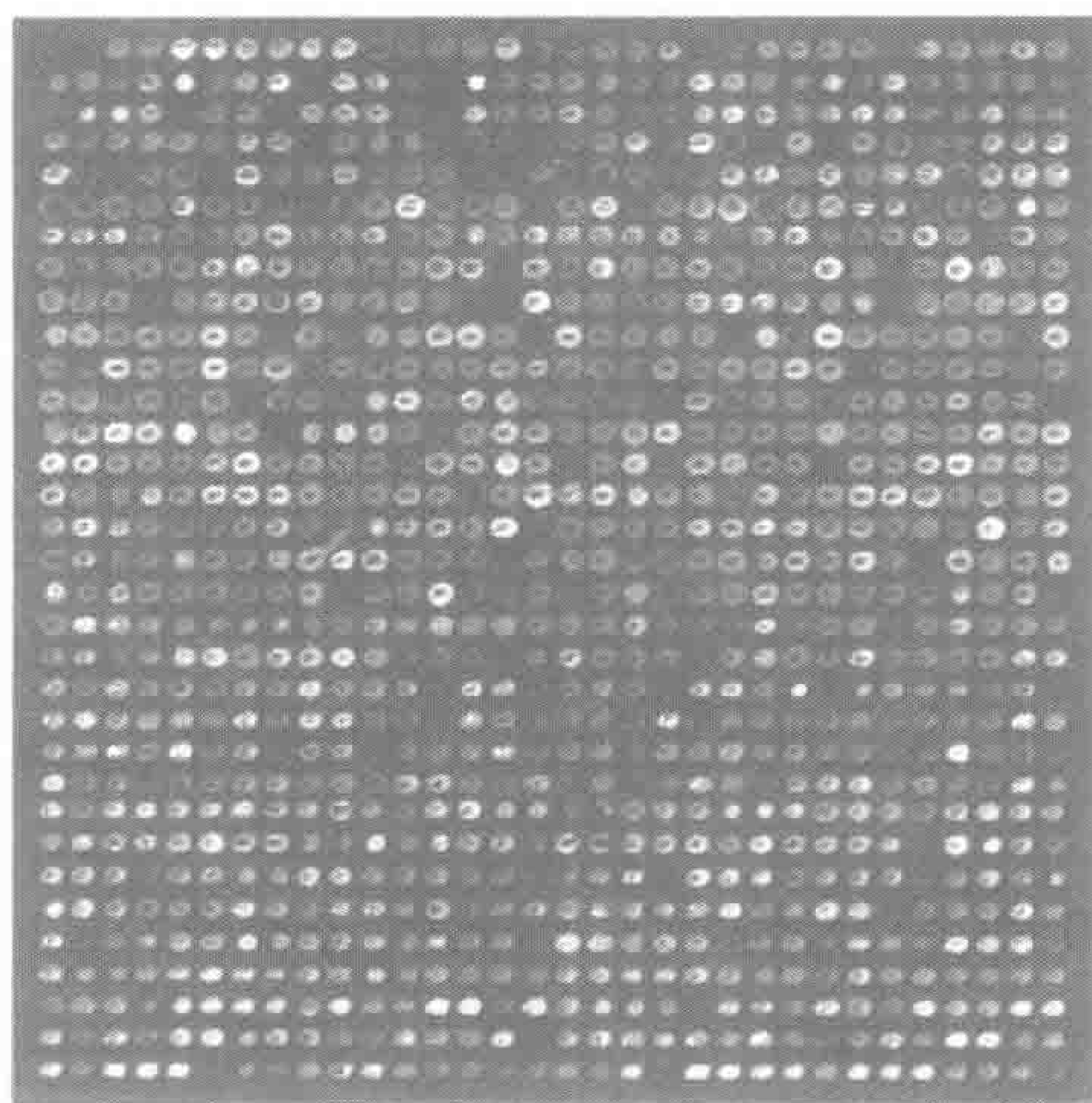


图 6.19　与一个 cDNA 克隆微阵列杂交获得的基因表达谱

一个 1.0cm^2 DNA 微阵列，含有人全血细胞 cDNA 文库的 1046 个 cDNA 克隆，与人类骨髓 mRNA 反转录制备的荧光标记 cDNA 探针杂交。荧光标记的共聚焦激光扫描以伪色的五颜六色的排列规模出现，来粗略限定表明表达水平：背景是紫色，然后进展为深蓝色、蓝绿色、黄色、橙色到红色（最丰富的表达）。芯片的平行格式保证了比较和差异表达精确的测量。图像由 Stanford 的 Mark Schena 博士友好提供，经 Nature Publishing Group 同意，再引自 Strachan 等（1997），*Nature Genetics* 16，126～132。

一旦构建或购买了微阵列，并且分离了将要研究的核酸源，就可以进行杂交反应了。靶序列用一个**荧光团**（fluorophore）标记，并使其与微阵列接触，能够形成探针靶序列异源双链，之后杂交冲洗使非特异结合标记最小化。大多数微阵列杂交使用两种荧光基团，通常为 Cy3（绿色通道激发）和 Cy5（红色通道激发）。

杂交后，应用高分辨**激光扫描仪**（laser scanner）检测结合的荧光标记，扫描过程涉及获取两种荧光基团的图像来构建成一个比例图像。利用**数字成像软件**（digital imaging software）分析阵列上每个点所发射的信号来获取最后的杂交模式。数字成像软件根据信号的强度，将其转换成调色板上的一种颜色（图 6.19）。

尽管建立 DNA 微阵列技术是最近发展起来的，但是它已经有了许多重要的应用，它对未来生物医学研究及诊断方法的影响将是深远的，主要的应用包括：

▶ **表达筛查**（expression screening）：大多数目前以微阵列为基础的研究焦点是 RNA 表达水平的监测（Granjeaud *et al.*，1999），它通过应用 cDNA 克隆微阵列（图 6.19）或基因特异性寡核苷酸微阵列（通常通过原位寡核苷酸合成构建，见图 17.18 和 19.6 以及节 19.3.3）来进行。

▶ **DNA 变异筛查**（DNA variation screening）：为了筛查 DNA 变异，需要应用寡核苷酸微阵列，并且目前已设计了几种并投入使用。通过使用 DNA 微阵列对人类线粒体基因组的再测序是此项技术用于评价个体中大规模序列变异的能力的一个成功验证（图 7.4）。对于已知疾病基因突变分析也具有巨大的潜力。另外，也有积极的努力以鉴定和分类人类单核苷酸多态性（SNP）标记。

（刘丽英　译）

进一步阅读

Molecular Probes. *Handbook of Fluorescent Probes and Research Products* at http://www.molecularprobes.com/handbook/

Sambrook J, Russell D (2001) *Molecular Cloning: A Laboratory Manual*, 3rd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schena M (1999) *Microarrays: A Practical Approach*. Oxford University Press, Oxford.

Schena M (2002) *Microarray analysis*. John Wiley and Sons, New York.

Various authors (1999) The Chipping Forecast. *Nature Genet.* **21** (Suppl.), 1–60.

Various authors (2002) The Chipping Forecast II. *Nature Genet.* **32** (Suppl.), 465–552.

参考文献

Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G (1999) Making and reading microarrays. *Nature Genet.* **21** (Suppl.), 15–19.

Feinberg AP, Vogelstein B (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132**, 6–13.

Granjeaud S, Bertucci F, Jordan BR (1999) Expression profiling: DNA arrays in many guises. *BioEssays* **21**, 781–790.

Hinds HL, Ashley CT, Sutcliffe JS *et al.* (1993) Tissue specific expression of FMR-1 provides evidence for a functional role in fragile X syndrome. *Nature Genet.* **3**, 36–43.

Kricka LJ (1992) *Nonisotopic DNA Probing Techniques*. Academic Press, San Diego, CA.

Ross MT, Stanton VPJ (1995) Screening large-insert libraries by hybridization. In: *Current Protocols in Human Genetics*, Vol. 1. (eds NJ Dracopoli, JL Haines, BR Korf, CC Morton, CE Seideman, DT Moir, D Smith). John Wiley & Sons, New York, pp. 5.6.1–5.6.30.

Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.* **16**, 301–306.

Schwartz DC, Cantor CR (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**, 67–75.

Strachan T, Abitbol M, Davidson D, Beckmann JS (1997) A new dimension for the Human Genome Project: towards comprehensive expression maps. *Nature Genet.* **16**, 126–132.

Wilkinson D (1998) In Situ *Hybridization: A Practical Approach*. 2nd Edn. IRL Press, Oxford.

第7章　DNA与基因结构、变异及表达的分析

本章内容

7.1　DNA测序与基因型分型

一个物种内的个体之间存在遗传差异，其中许多可能具有重要意义。因此，向基因组计划投入巨大努力固然重要，我们仍将继续关注定位和测定一个物种内，特别是我们人类这一物种内个体的DNA，并确定DNA变异的类型（**基因型分型**，genotyping）。

尽管用于大规模物理绘图和测序的传统方法仍有助于获取新的基因组序列，但它们将逐渐被更强有力的完全自动化直接方法、或通过扫描基因变异来明确DNA结构的方法所取代。例如，某段参考序列一经确定，基于寡核苷酸的微阵列杂交将能够实现对个体DNA的重测序。

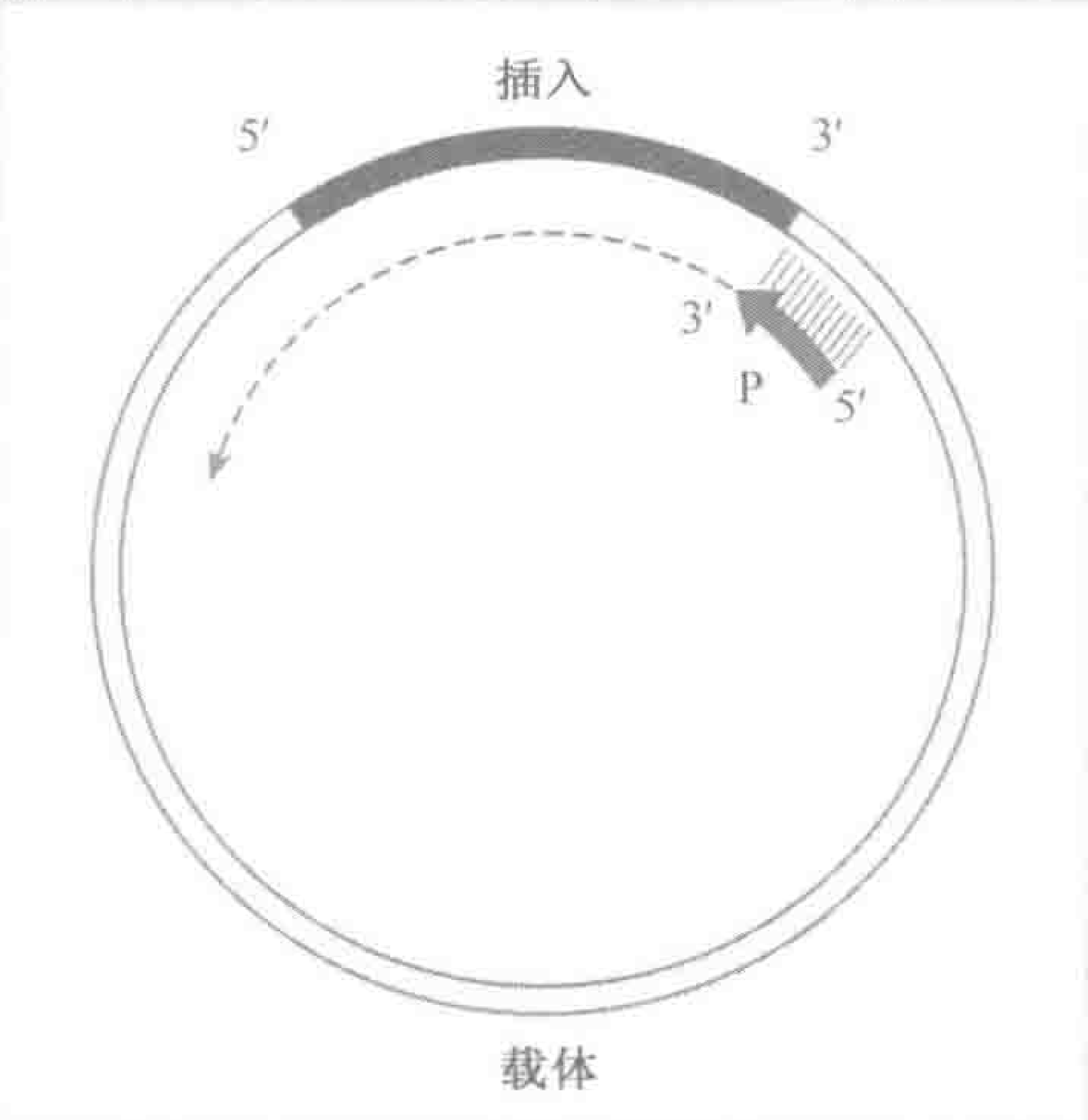
7.1.1　标准的DNA测序为使用碱基特异性双脱氧核苷酸链终止子的酶法进行DNA合成

DNA测序的化学方法（Maxam and Gilbert，1980）仍将继续有某些用途（如寡核苷酸测序，节7.2.4）。然而，目前绝大多数的DNA测序采用由Fred Sanger最先发展的酶法，这使得他赢得了第二个诺贝尔奖（前一个是因为发展了蛋白质测序）。在**双脱氧测序**（dideoxy sequencing）方法中DNA被制备成单链形式（框7.1）并作为模板在适当的DNA聚合酶作用下于体外产生新的互补DNA链。该方法包括四个平行的反应，每个反应中有四种dNTP以及一小部分作为碱基特异性链终止子的四种类似的**双脱氧核苷酸**（dideoxynucleotide，ddNTP）中的一种。

框 7.1　制备单链 DNA 测序模板

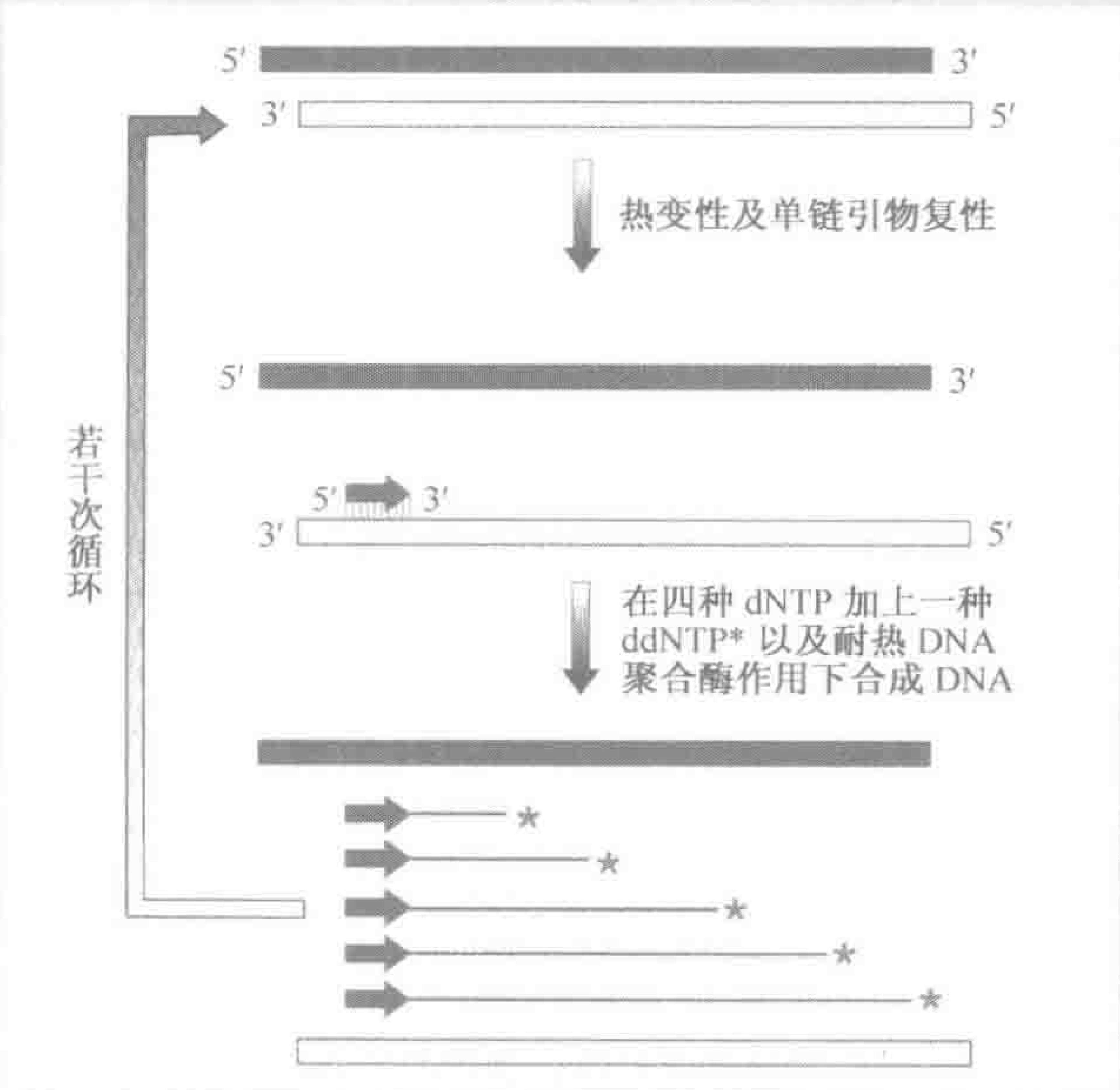
用于 DNA 测序的单链模板可来自：

▶ 用特化的克隆载体如 M13 或噬菌体制备单链 DNA 重组是一个应用广泛的方法——节 5.5.1 及图 5.18。本方法中互补 DNA 合成由一个**通用测序引物**（universal sequencing primer）引导，该引物与克隆位点旁侧的载体序列互补，因此可用于引导由该载体所制备的任何单链重组体的新链合成（见下图）；



一个通用测序引物可用于许多不同模板 DNA 的测序

▶ **循环测序**（cycle sequencing，亦称**线性扩增测序**，linear amplification sequencing）。与标准 PCR 反应类似，循环测序使用一种耐热 DNA 聚合酶和一个变性、复性及 DNA 合成的温度循环模式。不同的是循环测序仅使用一个引物并在反应中加入了一种 ddNTP 链终止子。仅使用一个引物使其产物以线性方式积累，而不像标准 PCR 反应中产物以指数性递增（见下图）。



循环测序涉及使用单一引物启动 DNA 合成的线性扩增

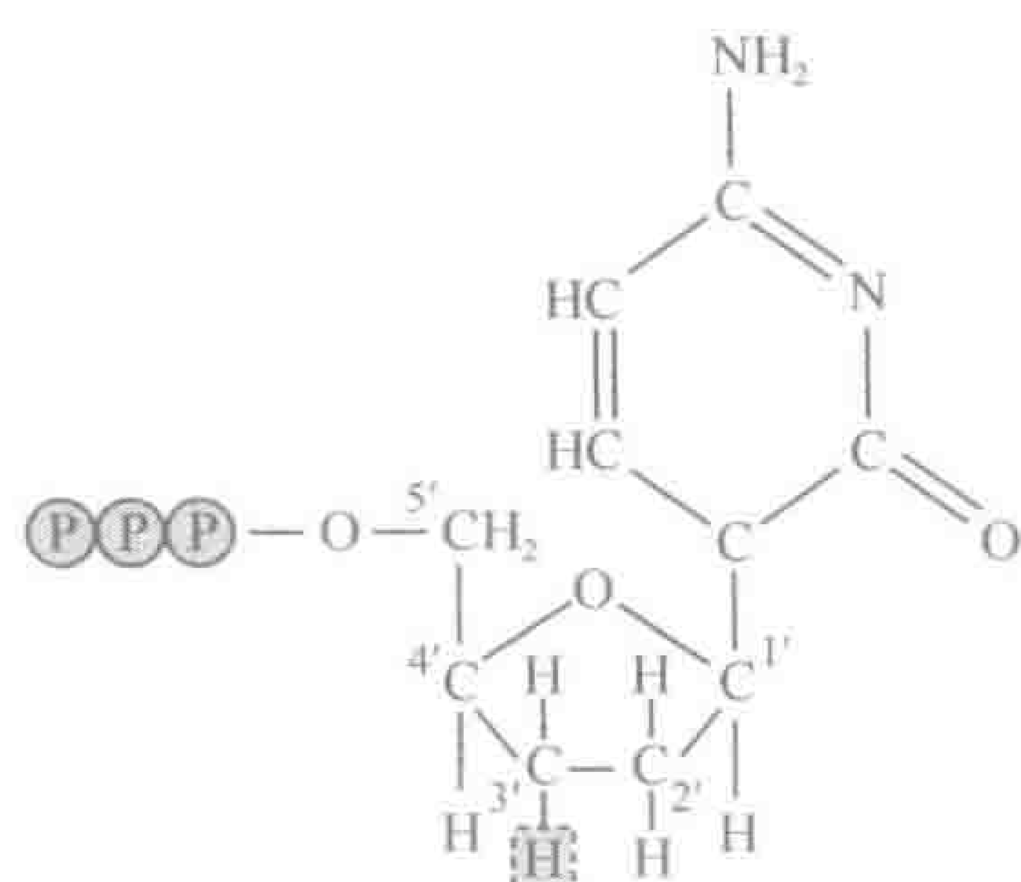


图 7.1 双脱氧核苷酸的结构，2'，3'双脱氧 CTP（ddCTP）

注：结合在正常核苷酸 3'碳上的羟基被氢原子替代（以阴影表示）。

ddNTP 同正常的 dNTP 非常相似；它们仅区别于前者在 3'碳以及 2'碳的位置上缺少一个羟基（图 7.1）。一个双脱氧核苷酸能通过在其 5'碳原子和前一个核苷酸的 3'碳之间形成磷酸二酯键而加入到延伸的 DNA 链中。然而，由于 ddNTP 缺乏 3'羟基，加入到一个延伸的 DNA 链的任何 ddNTP 均不能在其 3'碳原子上参与磷酸二酯键的形成，因而导致链合成的突然终止。

通过用某种特殊的放射性同位素基团或荧光素标记四种 dNTP 中的一种或引物，可使延伸的 DNA 链得以标记。通过将 ddNTP 的浓度设置为较其正常 dNTP 类似物低得多，可以使特异的 ddNTP 分子和过量的类似 dNTP 分子之间在加入延伸的 DNA 链时将存在竞争——如果加入一个 dNTP，DNA 链将继续延伸，但偶然加入一个 ddNTP 将导致链的终止。因此每个反应都是一个不完全的反应，因为链的终止会随机发生在任何一条 DNA 链对某种特殊类型碱基可能的选择上（图 7.2）。

因为 DNA 测序反应中的 DNA 通常为相同分子的一个群体，四个碱基特异性反应中的每一个将产生标记 DNA 片段的一个集合。四个反应中的每一个反应所合成的 DNA 片段具有一系列不同的大小。它们将具有一个共同的 5'端但不同的 3'端（5'端由测序引物确定；3'端将因为所选择 ddNTP 的插入随机发生于能接受特异性碱基的许多不同位置之上而不同）（图 7.2）。

大小相差甚至仅一个核苷酸的片段也能够在**变性聚丙烯酰胺凝胶**（denaturing polyacrylamide gel，一种含有高浓度变性剂——如 8M 尿素——的凝胶，前者能确保迁移的 DNA 保持单链状态）中依大小分离。在以前，DNA 测序反应使用放射性同位素 ^{35}S或^{32}P 标记，将干燥后的测序凝胶曝光于 X 光片后，通过跟踪自显影片上的连续条带人工读取序列。这种笨拙的方法现已被自动 DNA 测序方法取代。

7.1.2 自动 DNA 测序和基于微阵列的重测序

自动 DNA 测序

自动 DNA 测序（automated DNA sequencing）使用**荧光标记**（fluorescence labeling）：DNA 通过掺入带有**荧光团**（fluorophore，一种能发出荧光的化学基团，节 6.1.2）的引物或 dNTP 而标记。在四个碱基特异性反应中使用不同的荧光素使之不同于传统的 DNA 测序，全部四个反应能加样于一个泳道中。在电泳过程中，当 DNA 通过凝胶上一个固定点时，一个监测器将检测并记录信息（图 7.3A）。这使得结果以不同颜色荧光素的强度特征形式输出，同时以电子化形式储存信息（图 7.3B）。如果已知 DNA 序列为编码序列，输出结果可立即用不同的可读框翻译以推测多肽序列。

单链模板 DNA

+20　+1

5′ CGAATGCTCAGGCCATCATC 3′

5′

新 DNA 合成

引物

n 核苷酸长

依大小分离

A 反应（因 ddATP 而终止）　合成 DNA 的长度

AG　n+2
AGTAG　n+5
AGTCCGGTAGTAG　n+13
ACGAGTCCGGTAGTAG　n+16

C 反应（因 ddCTP 而终止）

CGGTAGTAG　n+9
CCGGTAGTAG　n+10
CGAGTCCGGTAGTAG　n+15
CTTACGAGTCCGGTAGTAG　n+19

G 反应（因 ddGTP 而终止）

G　n+1
GTAG　n+4
GTAGTAG　n+7
GGTAGTAG　n+8
GTCCGGTAGTAG　n+12
GAGTCCGGTAGTAG　n+14
GCTTACGAGTCCGGTAGTAG　n+20

T 反应（因 ddTTP 而终止）

TAG　n+3
TAGTAG　n+6
TCCGGTAGTAG　n+11
TACGAGTCCGGTAGTAG　n+17
TTACGAGTCCGGTAGTAG　n+18

n+20=G
n+19=C
n+18=T
n+17=T
n+16=A
n+15=C
n+14=G
n+13=A
n+12=G
n+11=T
n+10=C
n+9=C
n+8=G
n+7=G
n+6=T
n+5=A
n+4=G
n+3=T
n+2=A
n+1=G

图 7.2　双脱氧 DNA 测序依赖于体外 DNA 合成过程中碱基特异性链终止子的随机加入

测序引物特异性地结合至目的 DNA 序列的 3′区域，以所示方向引导一条互补 DNA 链的合成。四个碱基特异性反应平行进行，分别含有全部四种 dNTP 和一种 ddNTP。ddNTP 与正常的 dNTP 类似物在加入延伸 DNA 链中的竞争将产生一组不同长度的片段，这取决于双脱氧核苷酸（以上方的实心圆圈表示）在何处被插入。这些片段将具有共同的 5′端（由测序引物决定）及不同的 3′端。例如，在 A 特异性反应链中，延伸将持续至一个 ddA 核苷酸（以上方带实心红圈的 A 表示）被加入。这将产生一组长度分别为 $n+2$、$n+5$、$n+13$、$n+16$ 个核苷酸的 DNA 片段，如图右侧所示，若将这些片段按长度分离，将产生如下序列，从 $n=1$ 到 $n=20$ 为 GATGATGGCCTGAGCATTCG，即左上所示序列的反向互补序列。

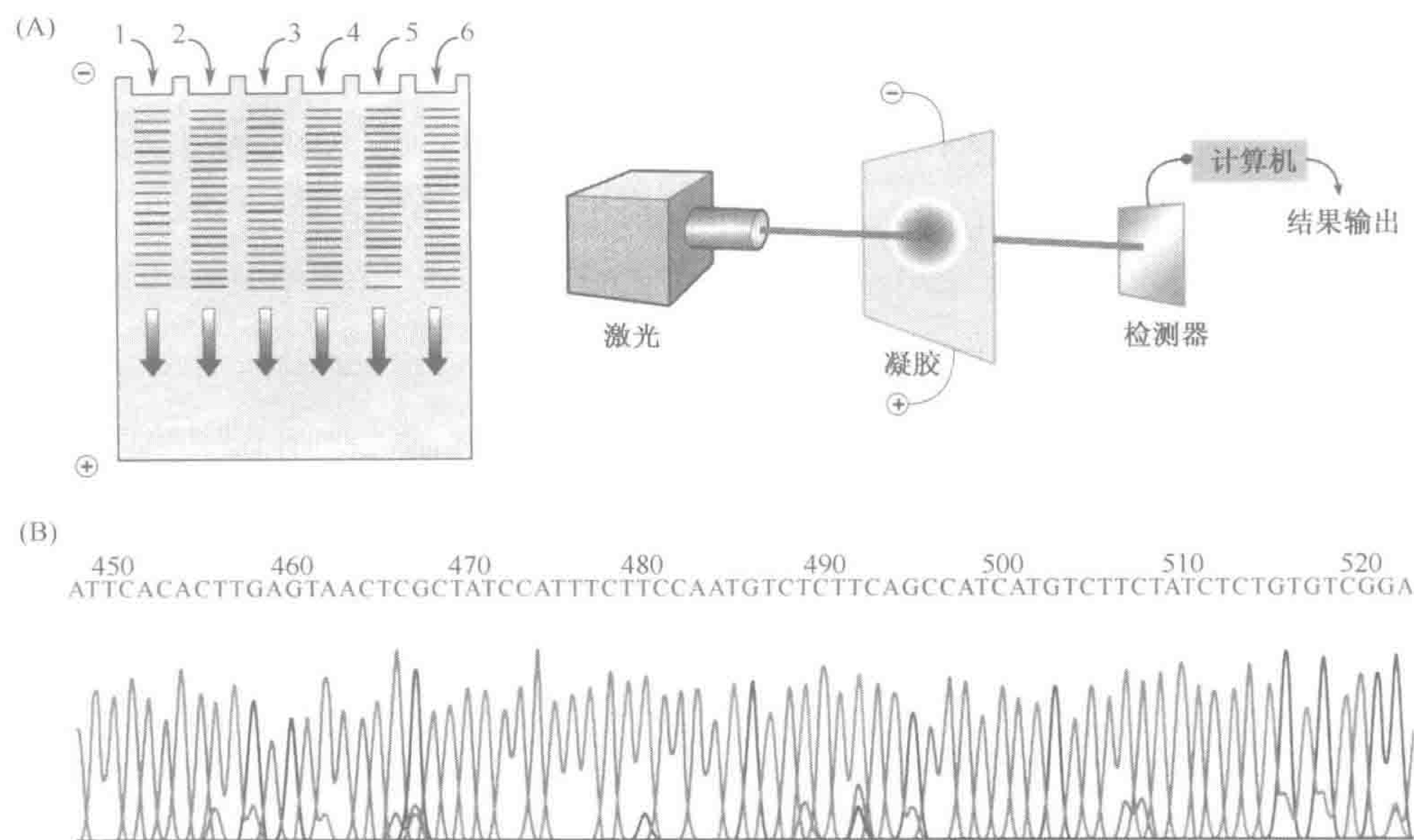


图 7.3　使用荧光引物的自动 DNA 测序

(A) **自动 DNA 测序的原理**。全部四种反应产物加样于电泳凝胶的一条泳道或单一凝胶毛细管中。四种不同的荧光染料在碱基特异性反应中被用作标记物（标记物可以通过结合至一个碱基特异性 ddNTP，或者通过结合在与四个反应对应的四套引物上而被加入）。在电泳过程中，激光束聚焦在凝胶上一个特定不变的位置。当单个 DNA 片段迁移通过这个位置时，激光将引起染料发出荧光。四种染料在不同波长下发出最强的荧光，其信息被以电子化形式记录下来，而解译出的序列被存入计算机数据库中。(B) **DNA 测序输出结果举例**。这里显示了一个由连续的染料特异性（因而是碱基特异性）强度特征表示的典型的序列数据输出结果。本例示意源自最近发现的人的多同源异型基因 PHC3 的一段 cDNA 序列（Tonkin *et al*.,2002）。数据由英国 Newcastle-upon-Tyne 大学 Human Genetics 学院 Emma Tonkin 博士提供。

尽管平板丙烯酰胺凝胶电泳被应用于许多自动 DNA 测序仪，高通量 DNA 测序目前常使用**毛细管测序仪**（capillary sequencer），这里 DNA 样品迁移通过充满凝胶的非常细长的玻璃毛细管（Meldrum，2000）。由于无需灌制大的凝胶，可实现更高程度的自动化。

应用微阵列杂交进行重测序

序列一旦被确定，原则上可以用它们作为参考序列来辅助设计**基于杂交的 DNA 测序**（hybridization-based DNA sequencing）的寡核苷酸。该方法需要在一块玻璃表面上原位合成寡核苷酸作为与待测序的标本 DNA 杂交的探针（寡核苷酸微阵列的原理见节 6.4.3）。通过微阵列杂交的重测序能以一种高度自动化的方式实现，其通量远超过自动 DNA 测序。一个实例为 16.5 kb 的线粒体 DNA（mtDNA）的重测序（Chee *et al*.，1966，图 7.4），然而通过微阵列杂交进行很大规模的测序却是一个艰巨的技术挑战。

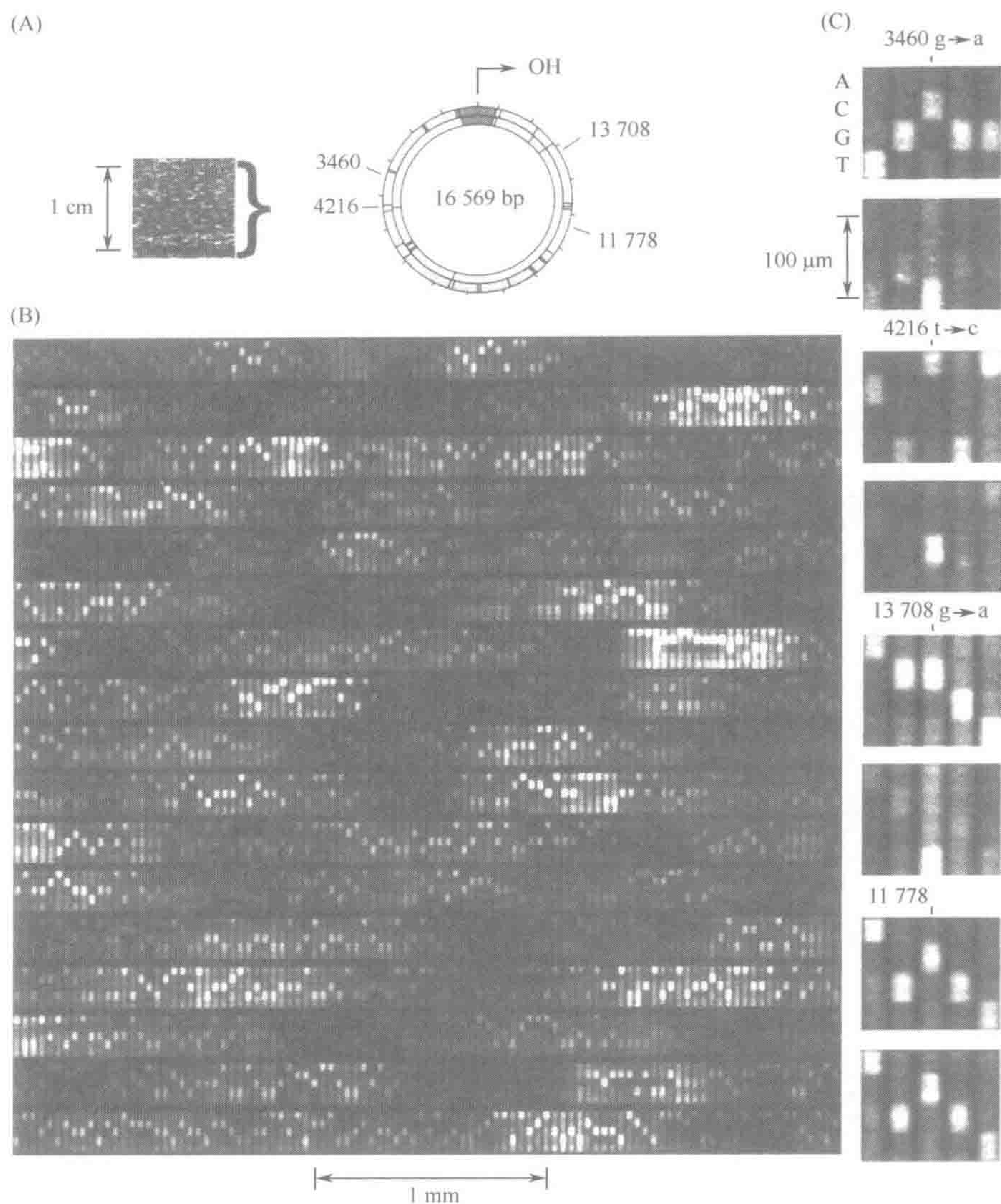


图 7.4 通过单芯片一寡核苷酸微阵列杂交重测序线粒体基因组

(A) 与 16.6 kb 线粒体靶 RNA（L 链）杂交的阵列图像以及 mtDNA 基因组图。(B) 放大后的杂交图的一部分。(C) 阵列在一个 16.6kb 样品中检测并阅读单碱基差异的能力如图示。两条不同靶序列并列杂交于不同的芯片，并比较了序列中四个不同位置的杂交模式。每对杂交图中的上图显示参考序列的杂交，下图显示来自一名 Leber 遗传性视神经病患者的标本所产生的图型。位于 3460、4216 和 13708 核苷酸位置的三个已知的病理突变被清晰检测。为了便于比较，图集中的第四排显示了 11778 位点周围的序列，其在两个样本中一致。经允许，重印自 Chee 等 (1996). Science 274，610～614，American Association for the Advancement of Science

7.1.3 限制位点多态性以及可变数目串联重复多态性的基本基因型分型

有两种基本类型的多态性适于使用简单的方法进行基因型分型：**限制位点多态性**（restriction site polymorphism，RSP）以及**可变数目串联重复**（VNTR）多态性。各种派生的多态性的描述见框 7.2。

框 7.2 适用于简单基因型分型方法的常见类型 DNA 多态性

正如将在第 11 章中详细阐述的，DNA 变异可发生在不同水平。偶尔能出现较大的改变，如数千甚至数百万碱基对的 DNA 片段的重复、插入、缺失和转座，其中有些与疾病相关，有些则不然。但 DNA 序列最为频繁的改变为单核苷酸置换、插入和缺失。它们通常与疾病无关，除非改变了某个编码序列或重要的调控序列。在人群中足够常见以至于不能用再发突变来解释的 DNA 变异被称为**多态性**（polymorphism，节 11.1）。可以进行简单基因型分型的多态性分为两种基本类型：

▶ SNP（**单核苷酸多态性**，single nucleotide polymorphism）。SNP 是由单个核苷酸改变而造成的多态性，常通过 DNA 测序（节 7.1.1；节 7.1.2）或引物延伸实验分析（框 18.2）；

▶ VNTR（**可变数目串联重复**，variable number of tandem repeat）多态性。由一组串联排列重复的不稳定重复单位数目改变而造成的一种多态性。这是一个涵盖了微卫星 VNTR 多态性和小卫星 VNTR（见下文）的总名称，但是通常被用来泛指小卫星类。

上述类型的多态性派生出的一些亚型为：

▶ RSP（**限制位点多态性**，restriction site polymorphism）。RSP 是 SNP 的一个亚类，其中核苷酸改变造成了一个限制位点的缺失或获得。这可以通过 PCR 或 Southern 杂交分型，这种情况下的多态性也称作 RFLP（见下文）；

▶ RFLP（**限制性片段长度多态性**，restriction fragment length polymorphism）。一种造成限制片段长度改变的多态性，用 Southern 杂交分型。可由两种方式产生：

● 作为一个 RSP 的结果（见上文）；

● 作为一个含有中等长度 VNTR 排列串联重复的限制片段长度变异的结果。旁侧的限制位点不变，但它们之间的长度可根据重复单位的数目扩大或缩小。

▶ **微卫星 VNTR** [microsatellite VNTR，又称**短串联重复多态性**（short tandem repeat polymorphism，STRP）或**简单序列重复多态性**（simple sequence repeat polymorphism，SSRP）]。VNTR 的一个类型。其序列较短（通常少于 100 bp）且重复单位较小，通常为 1～4 核苷酸——另见节 9.4.3；

▶ **小卫星 VNTR**（minisatellite VNTR，常易混淆地简写为 VNTR）。其序列大小为中等长度，重复单位通常为 9～65 bp——另见节 9.4.2。

RSP 基因型分型

单核苷酸多态性（SNP）常被用作研究常见病易感性的标记，多种自动化的高通量基因型分型方法已被设计用于检测这些多态性（框 18.2）。SNP 中的一部分可造成限制位点的丢失或获得（**限制位点多态性**，restriction site polymorphism，或 RSP），它们常用简单的基因型分型方法进行小规模分型。

过去，RSP 通过使用一个附近的探针进行 Southern 杂交用来检测限制性片段的改变分型，当采用这种分析方法时，RSP 被称作**限制性片段长度多态性**（restriction fragment length polymorphism，RFLP；基本原理见图 7.5A）。然而现在使用基于 PCR 的方法对一个 RSP 进行分型更为简便：根据多态性限制位点旁侧序列设计引物，扩增产物用合适的限制酶切割并经琼脂糖凝胶电泳依片段大小分离（图 7.6）。

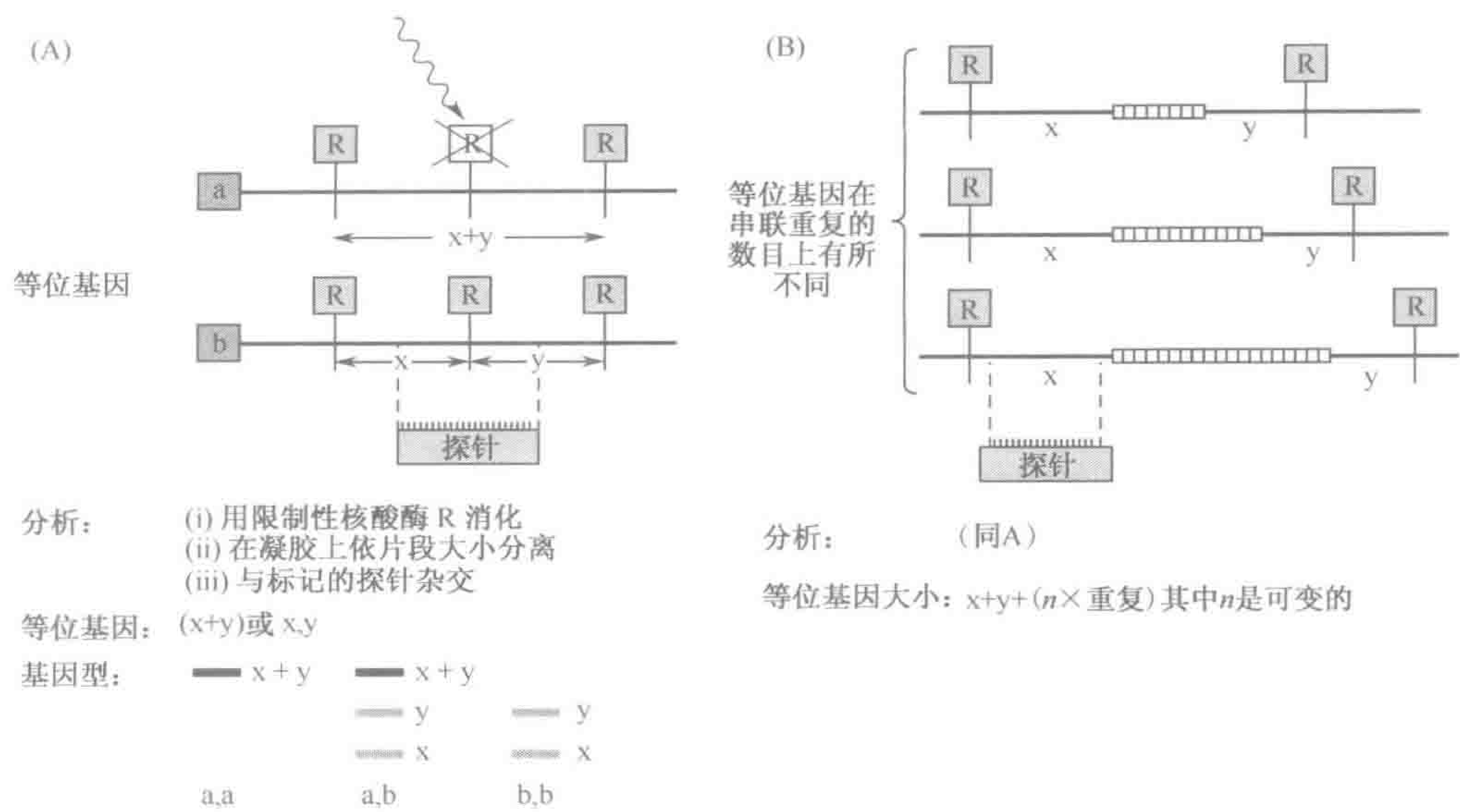


图 7.5　用基于杂交的分析对一个 RFLP 分型

(A) 对一个 RSP 型 RFLP 分型。这是 RFLP 的常见类型，由 DNA 序列的微小变化造成一个限制位点的丢失（或获得）所致。这类多态性更容易用 PCR 分析来分型（图 7.6）。(B) 对一个 VNTR 型 RFLP 分型。杂交分析仅用于涉及明显长度改变的 VNTR 的扩大或缩小，否则用 PCR 分析代替。

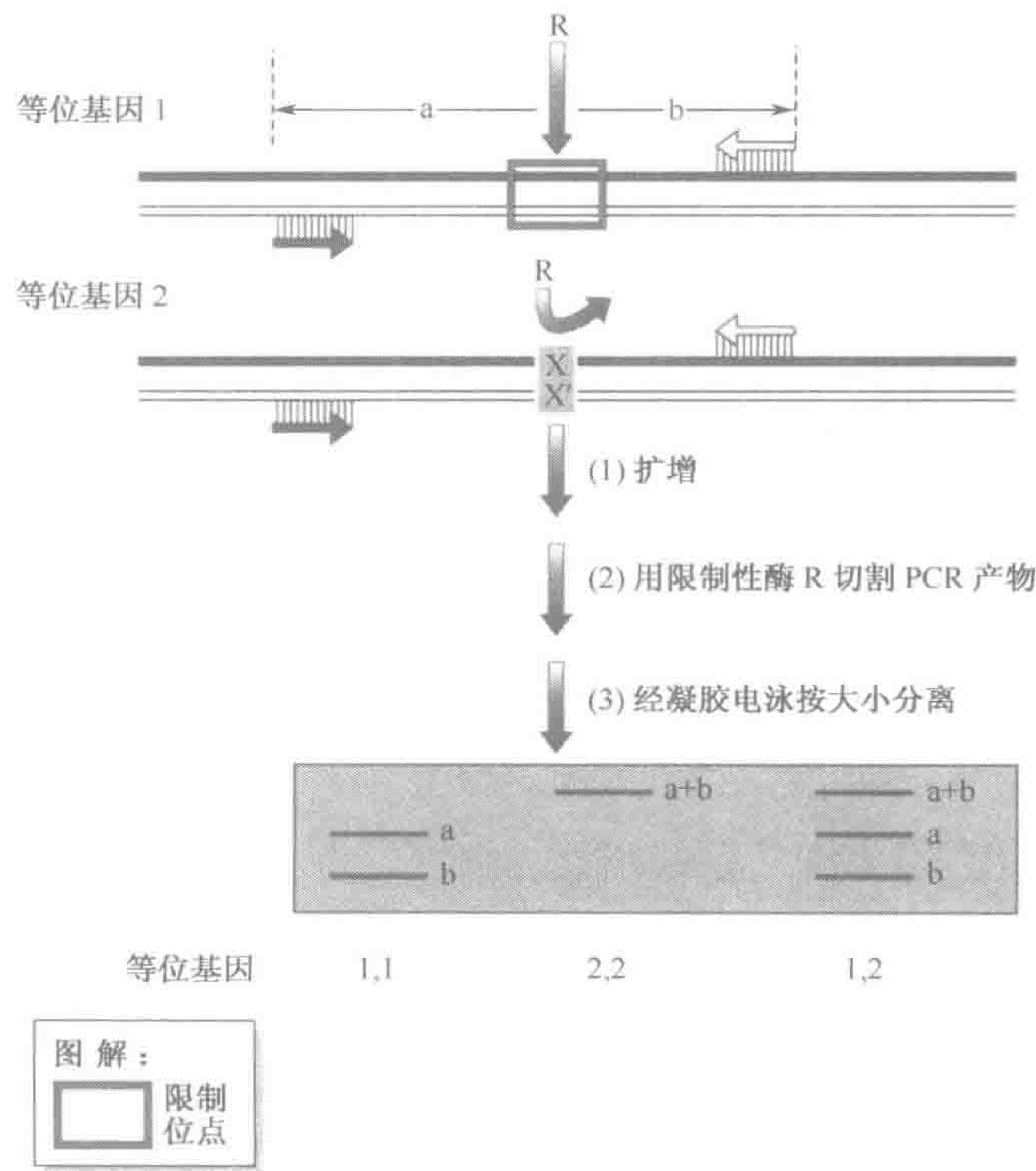


图 7.6　限制位点多态性可以容易地由 PCR 分型代替费力的 RFLP 分析

等位基因 1 和 2 区别在于一个多态，其改变了限制性核酸酶 R 的特异性限制位点的核苷酸序列。等位基因 1 具有这种位点，而等位基因 2 则具有改变了的核苷酸 X，X′，因而缺乏该位点。PCR 引物可以简单地由限制位点侧翼序列设计以产生一个短的产物。使用 R 酶消化 PCR 产物并按片段大小分离能对两个等位基因简便地分型。

VNTR 多态性的基因型分型

微卫星（microsatellite）多态性是 VNTR 的一种，其序列的大小及串联重复的大小比较短（其他名称见框 7.2），因此便于通过 PCR 进行分型。引物根据特定微卫星基因座旁侧的已知序列设计，使得 PCR 扩增整数倍重复单位的不同大小等位基因（图 7.7）。PCR 产物能进一步通过聚丙烯酰胺凝胶电泳分离大小。PCR 中通常包括放射性或荧光素核苷酸前体，其加入到小的 PCR 产物中有助于它们的检测。为了确保等位基因依大小分离，PCR 产物在电泳前变性。一个应用（CA）$_n$ 微卫星的例子见图 7.8。

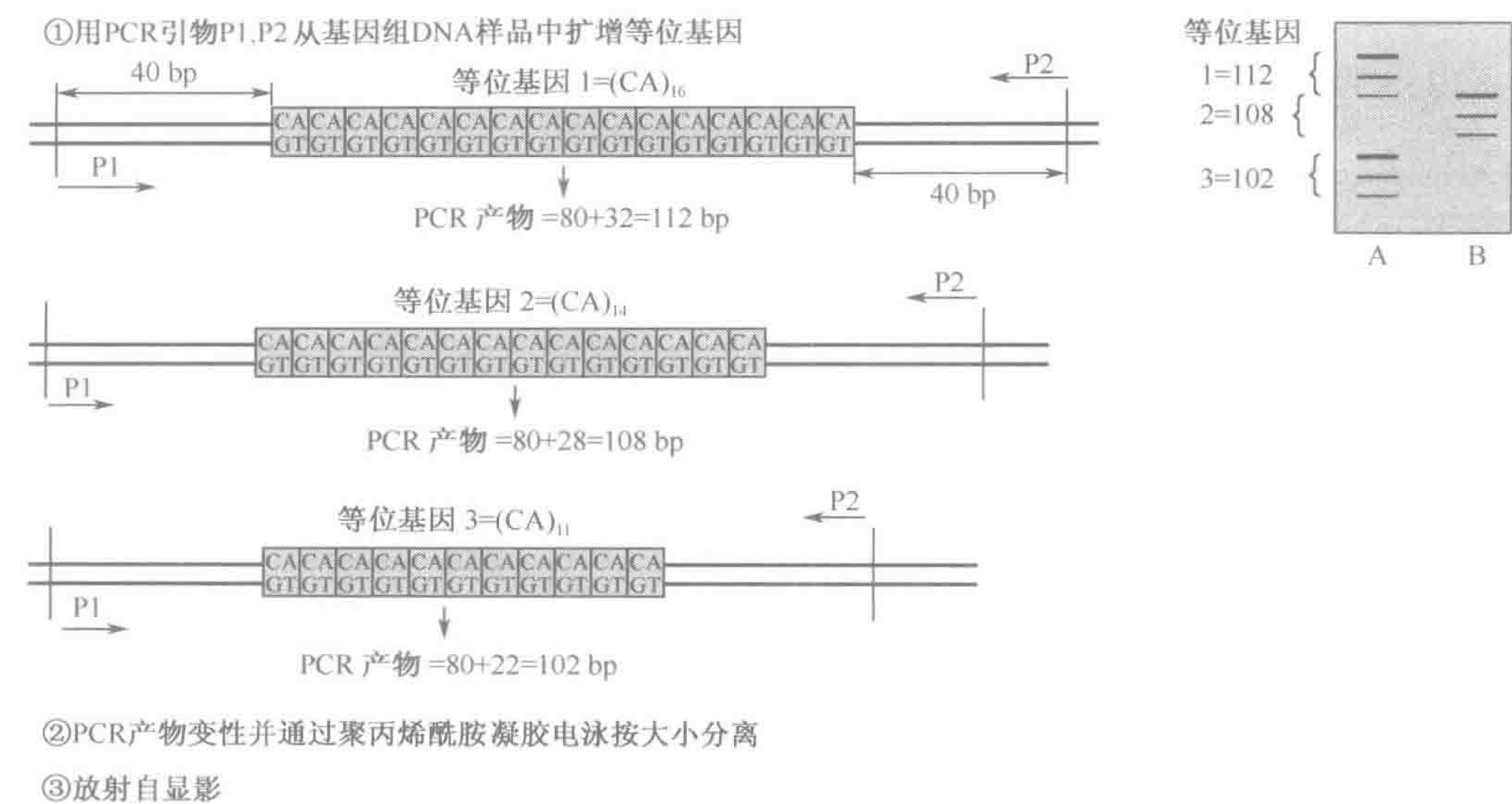


图 7.7　PCR 用于短串联重复多态（STRP，short tandem repeat polymorphism）的分型

本例示意对一个微卫星标记的分型，该标记为一个（CA）/（TG）双核苷酸重复多态性，因（CA）/（TG）重复数目变异而具有三个等位基因。在放射自显影片上各等位基因由一条上方的主带和两条次要的'阴影带'代表（图 7.8）。个体 A 和 B 的基因型（括号内）为：A（1，3）；B（2，2）。

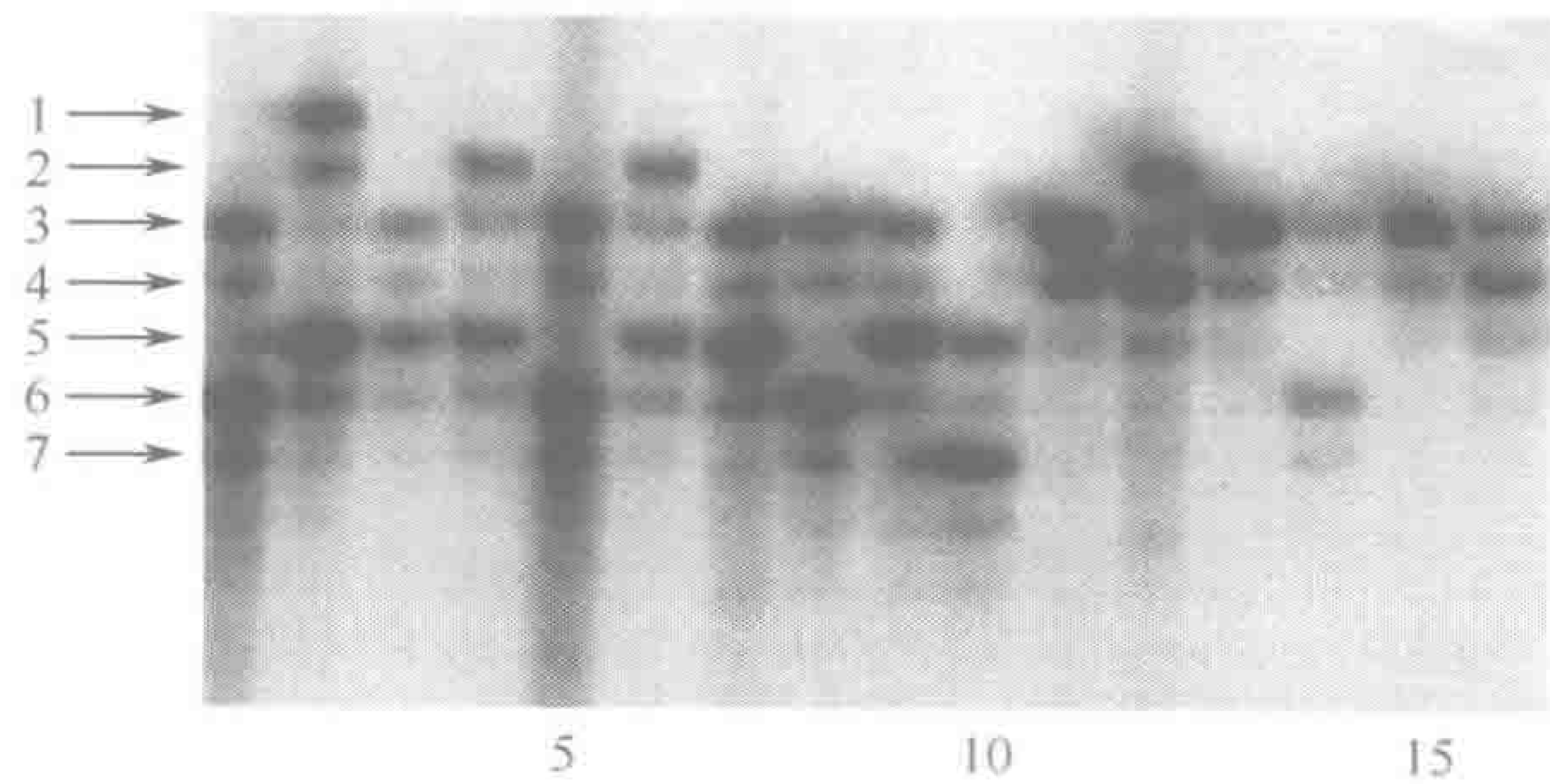


图 7.8　对 CA 重复分型的实例

本例示意用（CA）/（TG）标记物 D17S800 对一个大家系的成员进行分型。左侧箭头标示不同等位基因 1～7 的最强（主要）带。**注**：每个等位基因呈现一个强的上方带跟随着两个较低的阴影带，一个为中等强度位于紧靠强的上方带的下面，另一个非常微弱并位于紧邻第一条阴影带的下面。所示个体的基因型（括号中）如下：1（3，6）；2（1，5）；3（3，5）；4（2，5）；5（3，6）；6（2，5）；7（3，5）；8（3，6）；9（3，5）；10（5，7）；11（3，3）；12（2，4）；13（3，3）；14（3，6）；15（3，3）；16（3，4）。另外**注**：在后面几个例子里中间带特别强，因为它含有等位基因 4 的主带以及等位基因 3 的主要阴影带。滑移链错配（节 11.3.1）被认为是串联双核苷酸重复处产生阴影带的主要机制（Hauge and Litt，1993）。

微卫星 VNTR 多态性常用 Southern 杂交分析。DNA 被一种已知能剪切相关 VNTR 序列旁侧位点的限制性酶消化。这将产生一条含有 VNTR 以及附近特定序列 DNA 的限制片段。源自后一序列的探针能检测出作为一种 RFLP 的长度变化（图 7.5B）。

7.2 鉴定克隆 DNA 中的基因并确定其结构

在克隆的 DNA 中鉴定基因依赖于对基因特异性特征的分析。两个主要特点使基因的 DNA 区别于无编码功能的 DNA：（Ⅰ）整体上的高度进化保守，以及（Ⅱ）RNA 转录物的表达。在绝大多数情况下，RNA 转录物将进行剪接并翻译产生多肽（因此可通过具有长的**可读框**，open reading frame，ORF 来区分）。此外，脊椎动物基因常与 CpG 岛相关（框 9.3），这些特征使得可以用各种不同方法在克隆的脊椎动物 DNA 中鉴定基因（Monaco，1994）。

鉴定基因的常规方法

基因组 DNA 克隆一经获得，传统上首选简单的方法确定基因。为了检验是否存在表达，标准的方法包括筛查 cDNA 文库（节 5.3.5）、进行 RT-PCR（节 5.2.1）及将检测探针与 Northern 印迹进行杂交（图 6.13）。之后，表达分析常扩展为对组织切片中的 RNA 进行原位杂交分析（图 6.15）。寻找进化保守性的替代方法通常倚重于**动物印迹**（zooblot）来鉴定在一系列物种中强烈保守的序列（图 7.9）。最近，序列数据库同源搜索成为鉴定基因的重要方法：如果一段检测序列与某些其他编码序列非常相似，那么它很可能也是一段编码序列（框 7.3）。

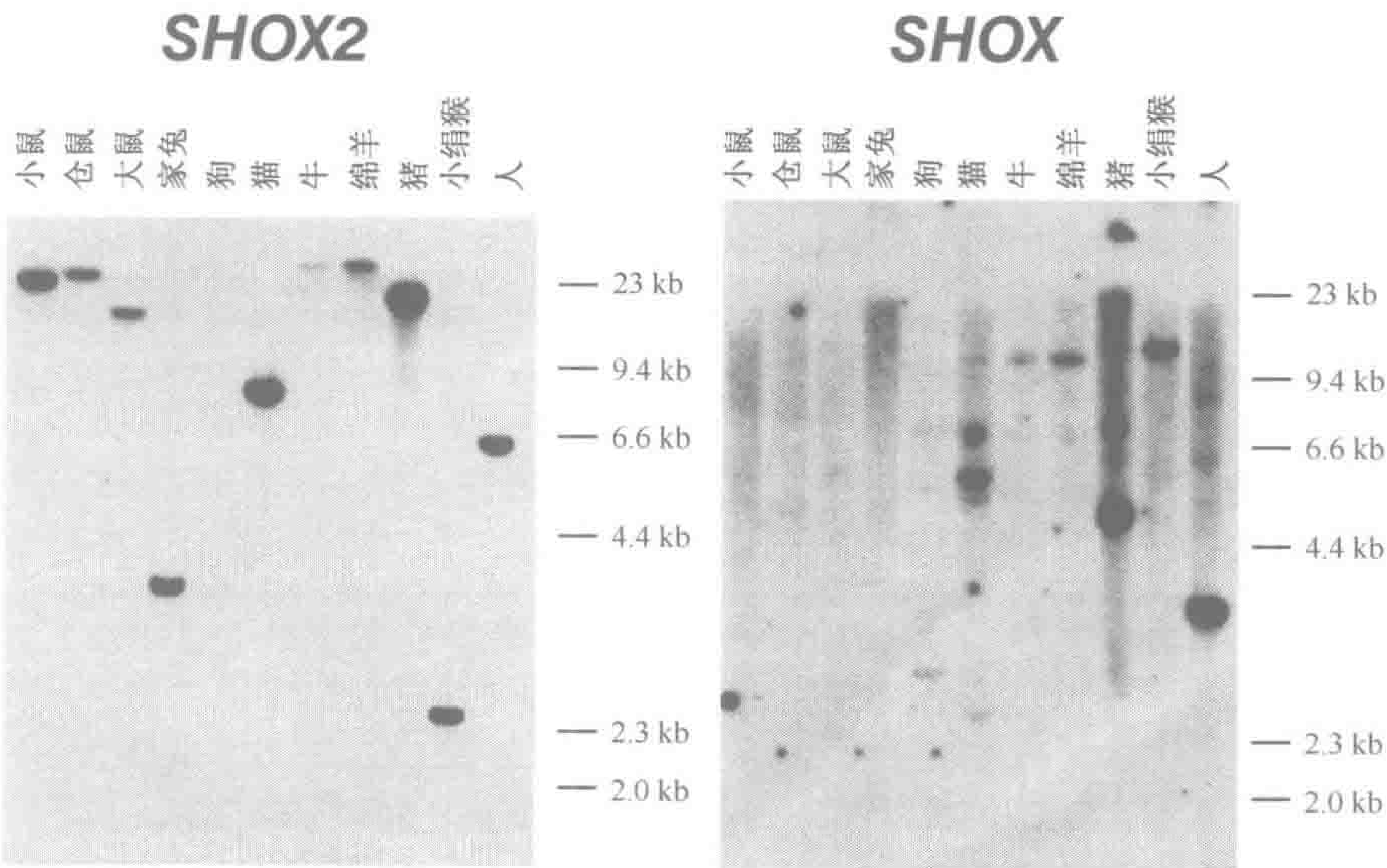


图 7.9 动物印迹杂交识别进化保守的序列

一些基因在物种之间显示显著的保守性；而包括本例所示在内的另一些则不然。人 *SHOX* 基因定位于 Xp/Yp 末端的**主要假常染色体区域**（major pseudoautosomal region）内（图 12.15）。它是一些以骨异常为特征的矮小综合征基因座，并可能是 Turner 综合征的一个重要因子。尽管在多种不同哺乳动物中保守，但作为进化中基因删除的结果，啮齿类动物却缺乏该基因（见右图左侧泳带），这很可能发生于啮齿类种系发生的早期阶段。最近对小鼠基因组的测序证实了 *SHOX* 同源基因的缺乏。相关的常染色体 *SHOX2* 基因保守的程度要高得多（左图）。经 Oxford University Press 允许，复制于 Clement-Jones 等（2000）. Human Molec. Genet. 9，696。

框 7.3　数据库同源性检索

使用强大的计算机软件，测试（查询）序列可用于搜寻序列数据库以发现显著相关的序列（**主体序列**，subject sequence）；继而比较查询和主体序列的最大序列相似程度（**序列同源性**，sequence homology）。普遍采用的是各种 BLAST 和 FASTA 程序（见 Ginsburg，1994 及下表）。

程序	比较
FASTA	一段核苷酸序列与另一段核苷酸序列相对比的数据库，或一段氨基酸序列与另一段蛋白序列相对比的数据库
TFASTA	一段氨基酸序列与以全部六种可读框翻译的核苷酸序列相对比的数据库
BLASTN	一段核苷酸序列与一段核苷酸序列相对比的数据库
BLASTX	一段以全部六种可读框翻译的核苷酸序列与一段蛋白序列相对比的数据库
EST BLAST	一段 cDNA/EST 序列与一段 cDNA/EST 序列相对比的数据库
BLASTP	一段氨基酸序列与一段蛋白序列相对比的数据库
TBLASTN	一段氨基酸序列与一个以六种可读框翻译的核苷酸序列相对比的数据库

注：因为诸如 FASTA 和 BLASTN 的比较程序在设计上有所不同，它们可能给出不同的结果（Ginsburg，1994）。所有以上程序均可以通过各种中心，如美国国家生物技术信息中心（http://www.ncbi.nih.gov）和欧洲生物信息学院（http://www.ebi.ac.uk）的互联网访问。

像 BLAST 和 FASTA 这样的程序用算法来寻找最佳的序列比对，并通常以测试序列（**查询序列**，query sequence）与程序在数据库中确定的各相关序列（**主体序列**，subject sequence）之间的一系列配对比较显示作为输出结果。

不同的方法可用来计算最佳序列比对。例如 Needleman 和 Wunsch（1970）设计的核苷酸序列比对算法寻求使匹配的核苷酸数目最大化。与之相反，其他程序如 Waterman 等（1976）的目标则是使错配数目最小化。当测试序列之间非常匹配并且具有相似，最好是相同长度时，序列比对的逐对比较相对简单。而当两个配对的序列彼此显著不同，尤其是因缺失/插入而长度明显不同时，计算最佳比对可能需要大量努力（见下面的表格）。

```
GATATTATCACTGGAGCCTGGCAGGAGCT        GATATTATCACTGGAGCCTGGCAGGAGCT
***  ****  ***************   *******    或    ***  ****  ***********   *  ********
GATTTTATGACTGGAGCCTGA-AGGAGCT        GATTTTATGACTGGAGCCT-GAAGGAGCT
```

序列共线性的难点。这里的两段核苷酸序列明显相关，但在上方显示的 GGC 序列处，对于与下方序列中的对应 GA 序列的最佳比对并不确定。

如果所研究的核苷酸序列是一段编码序列，那核苷酸序列比对就可能得利于通过平行比对编码序列的假定翻译可读框的氨基酸序列。这是因为存在 20 种不同的氨基酸但仅有四种不同的核苷酸。氨基酸序列的逐对比对也可以得利于对氨基酸化学亚类的考虑。**保守性替换**（conservative substitution）将导致氨基酸改变，但是新的氨基酸在化学上与被替换的氨基酸相近，且通常属于同一亚类（框 11.3）。因此，用于比较氨基酸序列的算法通常使用一种**矩阵评分法**（scoring matrix），其中成对的排列在一个 20×20 的矩阵中，较高的分值对应于相同和具有相似特性的氨基酸（如亮氨酸和异亮氨酸），较低的分值对应于具有不同特性的氨基酸（如异亮氨酸和天冬氨酸；见 Henikoff and Henikoff，1992）。典型的输出为两个序列亲缘关系百分比的综合结果，常称作%**序列一致性**（sequence identity，仅限于匹配的相同残基）和%**序列相似性**（sequence similarity，匹配的相同及化学性质相关的残基；见下面的表格）。

框 7.3 数据库同源性检索（续）

Score = 52.8 bits (125), Except = 9e-08

Identities = 39/120 (32%), Positives = 57/120 (47%), Gaps = 9/120 (7%)

```
Query: 1    AKLLIKHDSNIGIPDVEGKIPLHWAANHKDPSAVHTVRCILDAAPTESLLNWQDY-
            EGRTP 60
            A+LL++HD++   G PLH  A  +H  +    +  V+  +L+    W  Y  TP
Sbjct: 548  AELLLEHDAHPNAAGKNGLTPLHVAVHHNN---LDIVKLLLPRGGSPHSPAWNG
            Y---TP 601
Query:61     LHFAVADGNLTVVDVLTSY-ESCNITSYDNLFRTPLHWAALLGHAQIVHLLLER-
            NKSGTI 119
            LH A    +  V  L  Y  SN  S  +  TPLH AA  GH  ++V LLL  +  +G+
Sbjct: 602   LHIAAKQNQIEVARSLLQYGGSANAESVQGV--TPLHLAAQEGHTEMVALLL-
             SKQ-
             ANGNL 659
```

序列一致性与序列相似性。这里的 BLASTP 输出来自使用一个新鉴定的 inversin 蛋白氨基酸 165～283 作为查询序列检索 Swiss-prot 蛋白质数据库。这里显示的主体序列为小鼠胚胎细胞锚蛋白序列。程序不仅考虑**序列一致性**［sequence identity，120 个位置中的 39 个（32%）。在两条序列中具有相同残基；以红色字母标出］，而且还考虑**序列相似性**（sequence similarity，这里以‘正号’标出）其中 19 个位置具有化学性质相似的氨基酸（以+表示）。

除常规的基因识别方法外，还可应用两种更特异的方案：外显子捕获（一种人工 RNA 剪接分析）与 cDNA 选择。

7.2.1 外显子捕获通过应用一种人工 RNA 剪接分析来识别表达序列

RNA 剪接涉及 RNA 水平的外显子序列融合和内含子序列剔除。剪接体能通过识别位于外显子—内含子边界的特定序列在活体内实现这一过程。该特定序列包括：外显子与其下游（3′）内含子连接处的**剪接供体**（splice donor）序列，以及外显子与其上游（5′）内含子连接处的**剪接受体**（splice acceptor，见图 1.15）序列。黏粒和其他合适的基因组 DNA 克隆中含有内含子序列包围的内部外显子，因而含有功能性剪切供体和受体序列。

在克隆的基因组 DNA 中，外显子可以通过将 DNA 亚克隆入一个合适的表达载体并转染到适当的真核细胞株中而被识别，其中插入的 DNA 被转录为 RNA，而 RNA 转录物再进行 RNA 剪接。这种技术被称为**外显子捕获**（exon trapping，倘若利用 PCR 反应从一个经过剪接的 RNA 的 cDNA 拷贝中寻找外显子，常称为**外显子扩增**，exon amplification）。例如，在 Church 等（1994）的方法中，DNA 被亚克隆入一个表达载体 pSPL3（图 7.10A），该载体含有一个能在合适宿主细胞中表达的**人工小基因**（artificial minigene）。此小基因的组成为：猿病毒 40（SV40）基因组内含有复制起始点和一个强有力的启动子序列片段，被一个含有多克隆位点的内含子分开的两个有剪接能力的外显子以及一个 SV40 多聚腺苷酸化位点。

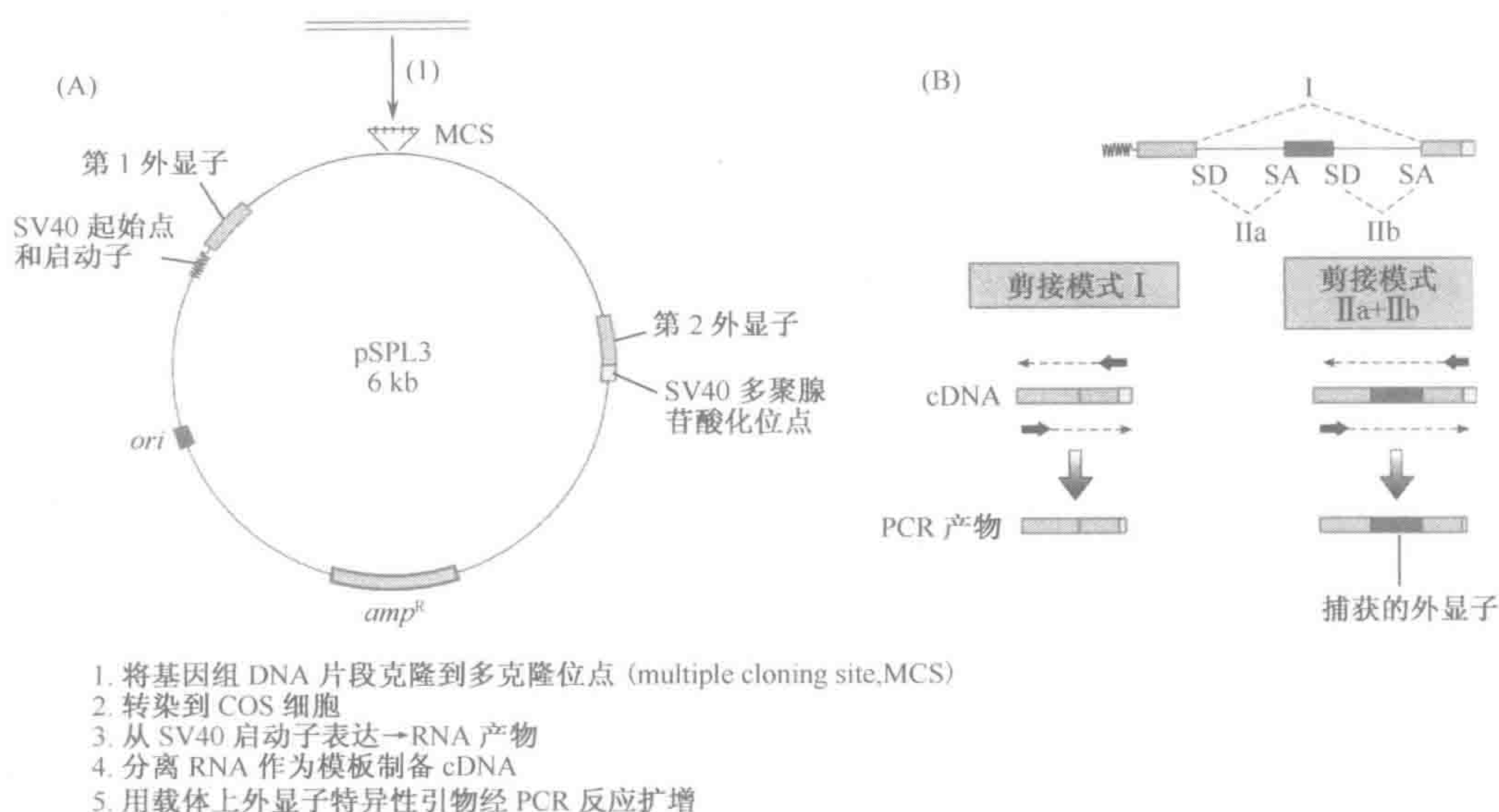


图 7.10　使用 pSPL3 载体进行外显子捕获

(A) **pSPL3 质粒载体**。这种穿梭载体能在 *E. coli*（利用 ori 复制起始点和对氨苄青霉素耐药性的选择）以及猴 COS 细胞（利用功能性 SV40 复制起始点）中增殖（Church *et al.*,1994）。pSPL3 质粒含有一个微小基因（黑色）；转录自 SV40 启动子开始，RNA 在宿主细胞的 RNA 剪接装置调控下进行剪接，导致两个载体外显子序列融合。(B) **剪接模式**。仅存在载体外显子的正常剪接模式以剪接模式 I 表示。如果克隆入 pSPL3 的基因组 DNA 片段包含一个带有功能性剪接供体（SD）和剪接受体（SA）序列的外显子，将出现一种不同的剪接模式（Ⅱa＋Ⅱb）。两种剪接模式通过使用各种载体特异性 PCR 引物在 cDNA 水平进行区分，在凝胶上依片段大小分离将使基因组 DNA 扩增的外显子得以回收。

重组 DNA 被转染至 COS 细胞中，正如在节 5.6.3 中所解释的，COS 细胞将使任何含有 SV40 复制起始点的环状 DNA 能够独立于细胞 DNA 进行复制。自 SV40 启动子的转录将产生一个常剪切为包含微小基因的两个外显子 RNA 转录物，然而倘若克隆入内含子的 DNA 片段含有一个功能性外显子，外源性外显子将被剪接至载体微小基因内的外显子上。用反转录酶产生一个 cDNA 拷贝后，使用载体外显子序列特异性引物的 PCR 反应能区分正常剪接和插入 DNA 中的外显子剪接（图 7.10B）。

7.2.2　cDNA 选择通过形成异源双链从基因组克隆中识别表达序列

cDNA 选择方法需要一个复杂克隆 DNA，诸如 YAC 插入片段，同一堆混杂的 cDNA，诸如一个 cDNA 文库中所有 cDNA 克隆中的插入片段，进行杂交（Lovett，1994）。该技术的原理是 YAC 中相应基因的同源 cDNA 优先与 YAC DNA 结合；多轮杂交将造成目的 cDNA 序列的富集从而鉴定相应基因。该方法需要大量用以封闭的重复 DNA 序列。

早期的方法使用经固定的 YAC，但现代的手段则采用一种液态杂交反应和生物素—链霉抗生物蛋白捕获方法（图 7.11）。与所有基于表达的系统相似，该方法依赖于适当水平的基因表达（在起始总体中同源 cDNA 不应太少）。此外，因为同源 cDNA 所

形成的异源双链可能不够稳定，含有短外显子的基因可能被遗漏。另一个问题是 cDNA 可能结合于与同源功能基因呈现高度同源性的假基因。

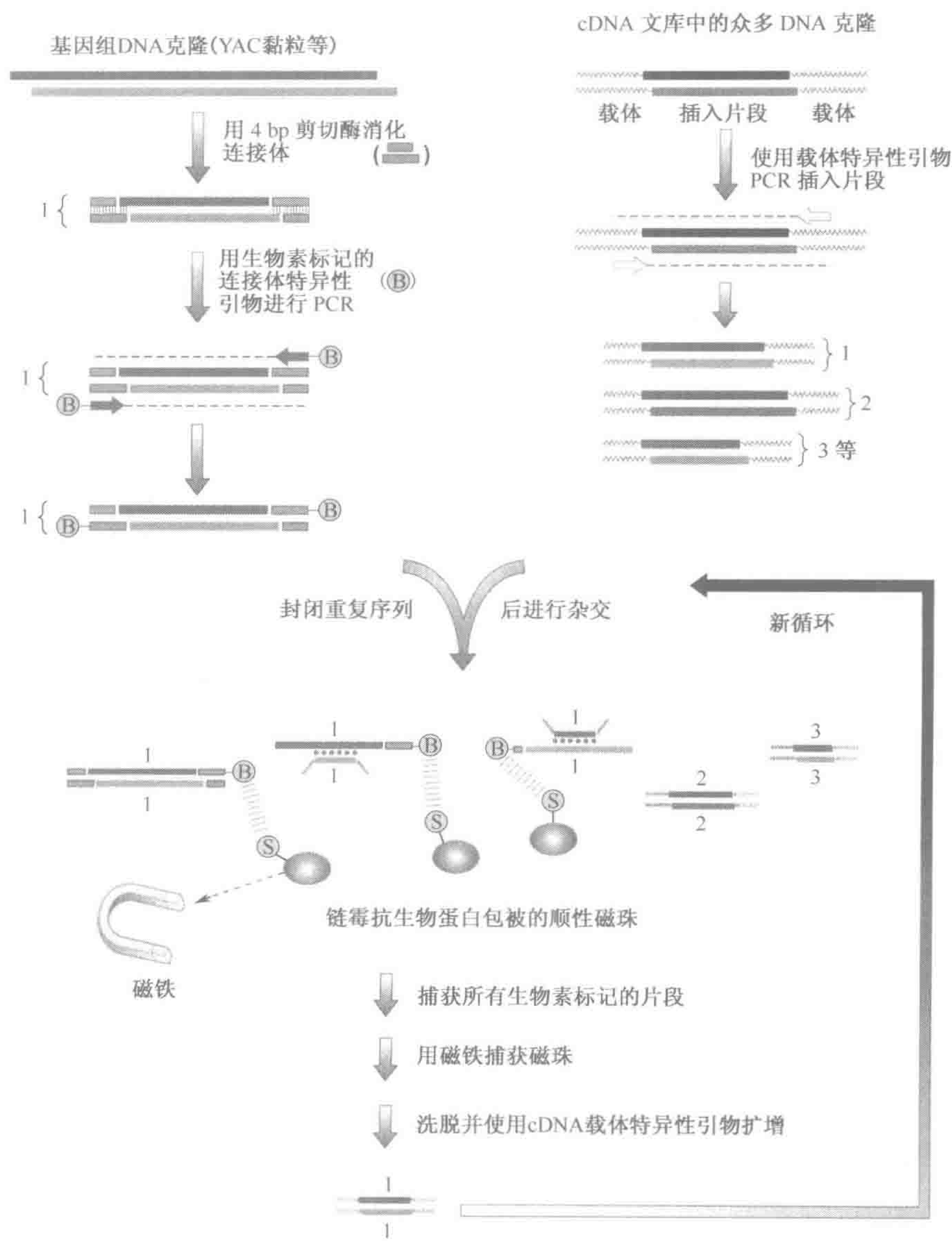


图 7.11　使用磁珠捕获筛选 cDNA

这种方法依赖于单一基因组 DNA 克隆（编号为 1）单链与一个混杂 cDNA 的群体，如某 cDNA 文库的插入片段（编号为 1，2，3 等），形成异源双链。基因组 DNA 链用一个生物素基因标记（与扩增过程中加入的 PCR 引物相连）。杂交反应有利于这些与基因组 DNA 克隆与同源的 cDNA 克隆间形成异源双链。本例中，基因组 DNA 克隆 1 与 cDNA 克隆 1 假设为同源，即含有共同序列，使相反的正义链结合在一起，形成一个异源双链。携带生物素基团的杂交产物（包括基因组 DNA－cDNA 异源双链）与链亲和素包被的顺性磁珠结合，通过磁场从其他成分中移出。分开的磁珠继而被处理洗脱含有生物素的分子，通过使用 cDNA 旁侧载体序列的特异性 PCR 引物，结合的 cDNA 被扩增。这一 cDNA 群体将经进一步杂交循环使目的 cDNA 富集。

7.2.3 获取全长 cDNA 序列：重叠克隆群与 RACE-PCR 扩增

确定基因结构最优先考虑的是获取一段全长 cDNA 序列，确定翻译起始和终止位点以及多聚腺苷酸化位点。

确定重叠克隆群

要得到一段全长 cDNA 序列，第一步就是筛查各种不同的 cDNA 文库，然后确定阳性克隆中插入片段之间重叠的程度（通过测序或基于 PCR/杂交的定位）。由此确立一系列重叠的 cDNA 克隆即一个 cDNA **克隆叠连群** [cDNA clone contig，例子见 Riordan 等（1989）中的囊性纤维化 cDNA 叠连群]。全长或选择性克隆测序能确定一致性序列，其可能提供全长 cDNA 序列。

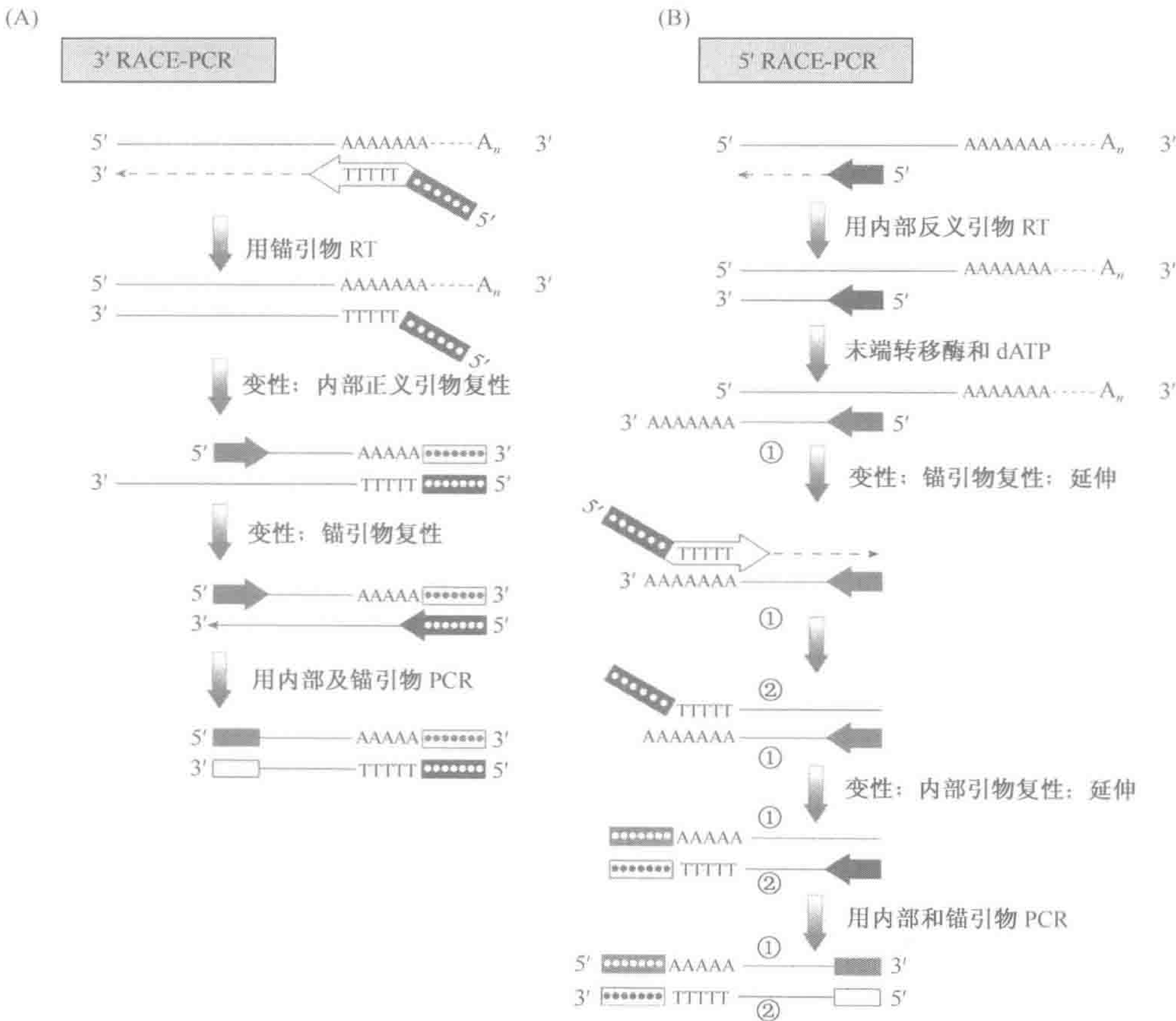


图 7.12　RACE-PCR 有助于自 cDNA 中分离 5′和 3′末端序列

RACE-PCR 的初始步骤为以 **5′追加诱变**（5′-add-on mutagenesis）的方式引入一个特异的末端序列（节 5.5.3）。(A) 3′ RACE-PCR。起始反义引物带有一个特殊的 5′延伸序列（锚序列，长度常超过 15 个核苷酸），该序列在反转录阶段被加入到 cDNA 转录物中。然后，一个内部正义引物用以产生一条短的次级链，其终止于与原始锚序列互补的序列。之后，PCR 由内部正义引物和一个锚序列引物启动。(B) 5′ RACE-PCR。这里一个内部反义引物用于引导自 mRNA 模板（链①）合成第一条 cDNA 链（链②）的一部分。一个 poly (dA) 用末端转移酶加入到 cDNA 的 3′端。第二条链的合成由具有一段特异性延伸（锚）序列的正义引物引导。这条链被用作内部引物进一步合成的模板以产生锚序列的互补拷贝。PCR 继而使用内部及锚序列引物完成。

RACE-PCR 延伸短 cDNA

RT-PCR 有助于确定转录物，但是 RT-PCR 的一个在延伸短 cDNA 序列以获得全长 cDNA 中尤其有效的变异形式为 RACE（cDNA **末端快速扩增**，rapid amplification of cDNA ends）技术（Frohman *et al*.,1988）。RACE-PCR 是 RT-PCR 的锚定 PCR 修饰。其原理为扩增 mRNA（cDNA）中一个事先鉴定的区域与一个偶联至 5′或 3′端**锚序列**（anchor sequence）之间的序列。其中一个引物根据已知的内部序列设计，另一个引物自相关的锚序列中选择（图 7.12）。

7.2.4 转录起始点定位与确定外显子-内含子边界

重要的调节序列通常位于转录起始点附近。尽管 5′RACE-PCR 能够寻找 mRNA 5′端对应的序列（因而是转录起始点），确定转录起始点优先采用的两种主要方法是：核酸酶 S1 保护和引物延伸。外显子-内含子结构能通过 cDNA 序列与同源的基因组 DNA 克隆序列对照而确定。之后，可以通过对启动子区域、5′和 3′旁侧序列以及内含子序列进行测序，从而在基因组水平完成基因鉴定。

核酸酶 S1 保护

S1 核酸内切酶是来自米曲霉（*Aspergillus oryzae*）的一种酶，可切割单链 RNA 和 DNA，但双链分子除外。定位一个基因的转录起始点，需要一个预计含有起始位点的基因组 DNA 克隆。DNA 克隆继而由适当的限制性内切核酸酶消化以产生一条预计含有转录起始点的片段。如图 7.13A 所示，与同源 mRNA 杂交并用 S1 核酸酶消化将确定转录起始点至限制片段未标记端之间的距离。如果需要更精确的定位，杂合双链中标记 DNA 片段可以通过 DNA 测序中的化学方法测序（Maxam and Gilbert，1980）。需要注意的是 S1 核酸酶定位亦可以一种非常相似的方式用于确定编码和非编码 DNA 之间的其他界限，如外显子-内含子边界（见下文）以及转录物 3′端。

引物延伸

这种方法与 S1 核酸酶保护相似。本方法中选取的限制片段必须比 mRNA 短，并且突出部分用反转录酶填平（图 7.13 B）。与 S1 核酸酶定位一样，通过使用 Maxam 和 Gilbert（1980）的化学测序方法来测序标记的 DNA 链能实现转录起始点更精确的定位。

确定外显子-内含子结构

并非所有的人类基因都有内含子（表 9.5），但只要存在，它们通常比外显子大。内含子的存在与否常能从同源基因组与 cDNA 克隆之间的比较来推测。一旦全长 cDNA 序列被确定，测序引物可根据需要自各种 cDNA 片段设计，并用于以变性的基因组克隆作为 DNA 测序模板的**循环 DNA 测序**（cycle DNA sequencing）（框 7.1）。所获得的序列应跨越一个外显子-内含子边界，除非外显子非常时可能需要额外的测序引物。当然，当一个基因组已被测序，仅需要将 cDNA 克隆与可用的基因组序列进行对比即可。

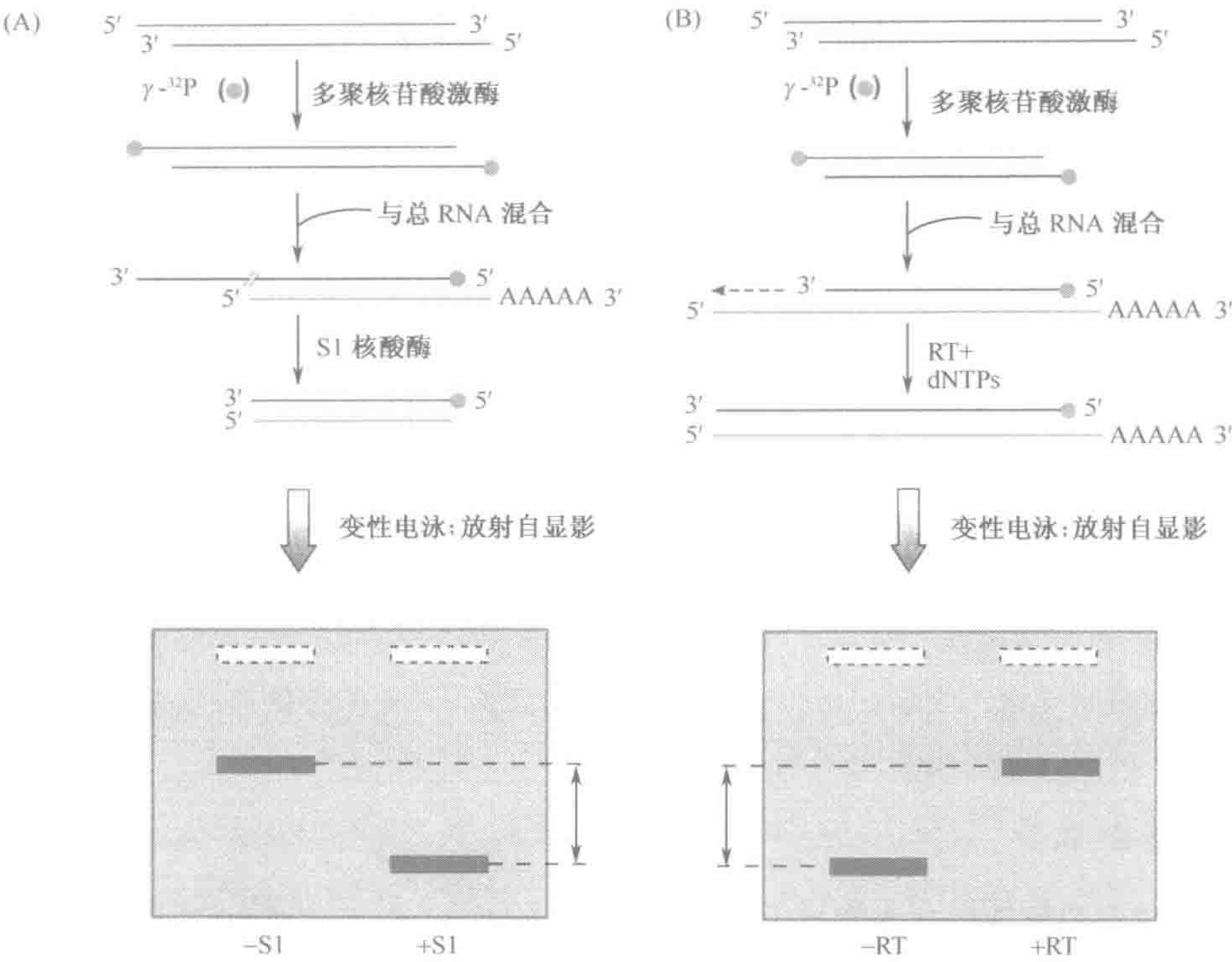


图 7.13　转录起始点可通过 S1 核酸酶保护或引物延伸分析定位

(A) S1 **核酸酶保护分析** (nuclease S1 protection assay)。来自克隆基因 5′端的一个限制片段预计含有转录起始点。该片段 5′端经末端标记后变性，并与可能表达相应基因的细胞总 RNA 混合。同源 mRNA 能与反义 DNA 链杂交形成 RNA-DNA 杂合双链。进一步用核酸酶 S1 处理将导致突出的 3′DNA 序列被不断切割直至与 mRNA 的 5′端杂交的 DNA 位点。在变性电泳凝胶上依片段大小分离将确定起始 DNA 和核酸酶 S1 处理后的 DNA 的长度差异。(B) **引物延伸分析** (primer extension assay)。本方法中选取的预计含有转录起始点的限制片段特意较短。与同源 mRNA 杂交将为 mRNA 留下一个突出的 5′端。DNA 可作为反转录 (RT) 的一条引物将其 3′端延伸至 mRNA 的 5′端。反转录处理后 (+RT) 较处理前 (−RT) 长度的增加能确定转录起始点。两种方法中更精确的定位可通过 S1 或 RT 处理后进行 DNA 测序来实现。

7.3　研究基因的表达

7.3.1　表达筛查的原理

表达筛查可以用不同的技术在不同水平进行。筛查的目标可以是 RNA 转录物或蛋白质。蛋白质表达通常用高度特异性抗体跟踪，而 RNA 转录物则可用几种不同类型的方法跟踪。这些方法常包括使用特异性反义核苷酸探针进行分子杂交，或者某些 RT-PCR 的变异形式。然而，另一些可选的设计精巧的方法已被发明，例如 SAGE (**基因表达系列分析**，serial analysis of gene expression)，其通过跟踪有代表性的短序列标签追踪大量单一转录物 (节 19.3.2) 基因表达中的重要参数为研究材料的来源、表达的分辨率和通量 (图 7.14)。

	分辨率	通量	例子
RNA	高	低	组织原位杂交 细胞原位杂交
	低	低	Northern 印迹杂交 RNA 斑点印迹杂交 核酸酶保护试验
	低	高	DNA 微阵列杂交 差异显示 基因表达系列分析（serial analysis of gene expression，SAGE）
蛋白质	高	低	免疫组化 免疫荧光显微镜
	低	低	免疫印迹（western 印迹）
	低	高	2-D 凝胶电泳 质谱

图 7.14 表达作图可以在不同水平进行

通量指能够同时被研究的基因/蛋白质数目。

研究材料的来源

用于研究的材料非常广泛。通常制备 RNA/cDNA 或蛋白粗提物，而在其他情况下，表达可自组织切片乃至固定以保持活体形态的整个胚胎中取材。此外表达也可以在组织培养的活细胞中被研究（但通常的问题是与活体中同种细胞相比它们的代表性如何）。在组织呈光学透明的活体实验生物中，基因表达可在荧光标签的辅助下被跟踪。

新型激光捕获显微切割（laser capture microdissection）法涉及使用激光来显微切割组织，以便从活检组织和染色组织甚至单个细胞等来源中获得纯细胞群（Schutze and Lahr，1998；Simone *et al*.,1998）。这方面进展使得基因表达分析集中于单个细胞，或比细胞株更能代表体内状态的单一细胞群。

基因表达的分辨率

一些方法仅被设计用来跟踪基因在 RNA 提取物或蛋白提取物中的总体表达。这类低表达模式通常作为实验的第一步。除能在不同组织中检测表达外，这些方法也能提供关于产物大小以及潜在异构体的信息。与之相比，高分辨表达可以通过跟踪基因在一个细胞或多组空间结构上可代表正常体内结构的细胞或组织中的表达模式而实现。

基因表达的通量

一些方法被设计每次仅获取一个或极小数目基因表达数据（低通量表达）。其他的方法则能每次同时跟踪许多基因的表达，在基因组计划已鉴定出某个有机体所有基因的

情况下能够进行**全基因组表达筛查**（whole genome expression screening，节 19.3）。

7.3.2 基于杂交的基因表达分析：从单个基因分析到全基因组表达筛查

传统的基于杂交的基因表达筛查通量较低，集中在一次分析来自一个或仅几个基因的 RNA 转录物，但分辨率可以从低到高有所不同。然而，最近基于微阵列的基因表达分析引出了一类新的通量非常高的低分辨率表达分析方法，能够一次同时检验来自上千种基因的 RNA 转录物。

Northern 印迹杂交

这种方法通过将基因或 cDNA 探针与由不同组织或细胞株制备的总 RNA 或 poly (A)$^{+}$ RNA 提取物杂交形成了低分辨表达模式。因为 RNA 在凝胶中以片段大小分离，能够估计转录物的大小。在一个泳道中出现多条杂交带可能提示存在长度不等的异构体（图 6.13）。

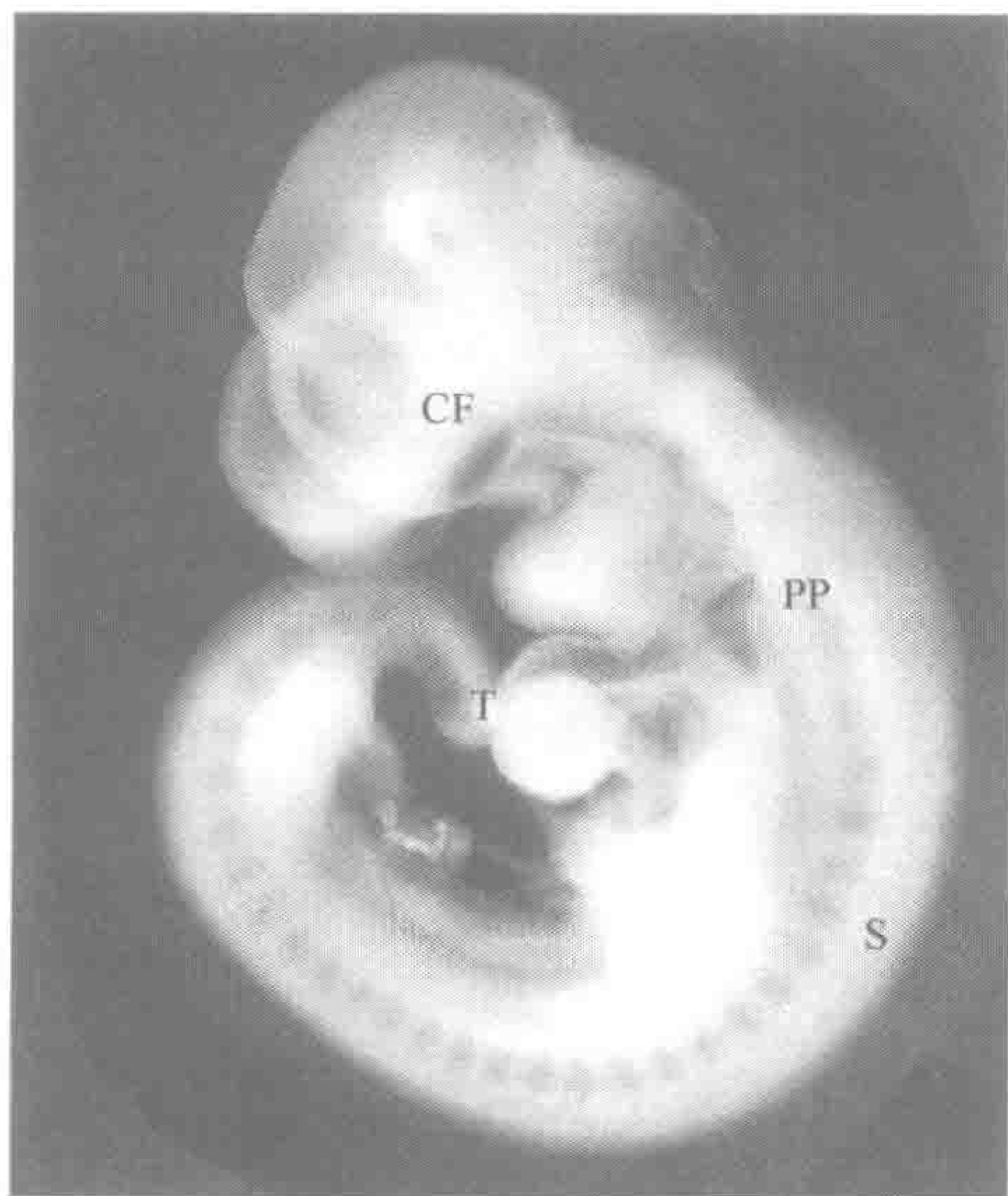


图 7.15 整装原位杂交
本例显示在一个 E9.5（胚胎为 9.5 天＝交配后 9.5 天）小鼠胚胎中 Pax9 的表达。表达在颅面区域（CF：将发育成为鼻间充质）、咽囊（PP）、体节（S）和尾芽（T）中较为明显。图像由英国 Newcastle-upon-Tyne 大学 Institute of Human Genetics 的 Heiko Peters 博士惠赠。

组织原位杂交

在不同组织与细胞群中 RNA 高分辨率的空间表达模式常通过**组织原位杂交**（tissue *in situ* hybridization）获得。通常，组织经冷冻或石蜡包埋后用切片机切削成非常薄的切片（如 5 μm 厚）放在显微镜载玻片上。将适当的基因特异性探针与载玻片上的组织杂交，可得到代表原始组织中 RNA 分布的详细表达图像（图 6.15）。包括胚胎组织的常用组织具有规格微小，能够在单张切片上筛查许多组织表达的优点。

整装原位杂交

组织原位杂交的拓展是在一个完整胚胎中研究表达。**整装原位**（whole mount *in situ*）杂交是在模式脊椎动物完整胚胎中跟踪发育过程中表达的常用方法。由于开展相应的人类基因分析具有的伦理方面和实际上的困难，因此很大程度上有赖于小鼠胚胎上进行的分析的推测（图 7.15）。相对大量的可用组织使得该方法相对灵敏并且该技术的自动化促进了其普及。

单细胞基因表达谱

使用适当标记的探针能在单一细胞中跟踪特异性 RNA 序列，以确定 RNA 加工、运输和细胞质定位。通过使用定量荧光原位杂交（FISH）和数字成像显微镜，甚至能

在原位观察单一的RNA转录物（Femino *et al.*,1998）。进一步的改进使用不同类型寡核苷酸探针组合，这些探针在多个位点用多种具有独特光谱的荧光素中的一种标记。这使得多个基因的转录物能够同时被跟踪（Levsky *et al.*,2002）。

用微阵列进行大规模表达筛查

基因表达筛查因为能够在玻璃表面制备高密度寡核苷酸或cDNA克隆**微阵列**（microarray）而改变（节6.4.3）。在一些已测序完全的基因组中，其提供了**全基因组表达筛查**（whole genome expression screening）的可能，借此一个生物体内每个基因的表达可以被同时检测（节19.3）。

7.3.3 基于PCR的基因表达分析：RT-PCR和mRNA差异显示

如在节5.2中所述，PCR的巨大优势在于快速、灵敏和简单。尽管并不适于提供表达的空间模式（以一种比如组织原位表达的方式），它能迅速提供可能有重要价值的总体表达模式。

常规RT-PCR

反转录酶PCR（RT-PCR：基本原理见节5.2.1）能提供一个特定基因的粗略表达量（用于细胞类型或组织不易大量获得的情况下，如植入前的早期人类胚胎，Daniels *et al.*,1995）。PCR的极其敏感使RT-PCR也能用于研究单一细胞内的表达。此外，RT-PCR可用于识别和研究一个RNA转录物的不同异构体。例如，不同的mRNA异构体可能通过选择性剪接产生，并在外显子特异性引物寻找预期产物及额外扩增产物时被识别（Pykett *et al.*,1994）。

mRNA差异显示

通过使用部分简并PCR引物（引物在某些位置的碱基特意灵活选取）能设计改进形式的RT-PCR，同时跟踪多个基因的表达。此类技术中的一种是**mRNA差异显示**（mRNA differential display）。它使用一种改进的oligo（dT）引物，其3′端具有一个不同的单核苷酸（即A、C或G，而不是T）或一个不同的双核苷酸（如CA）。因此，它将与部分mRNA的poly（A）尾结合（Liang *et al.*,1993）。例如，若用寡核苷酸TTTTTTTTTTT**CA**（$=T_{11}$ CA）作引物，它将优先引导那些双核苷酸**TG**位于poly（A)尾之前的mRNA合成cDNA。

上游引物通常是随机的短序列（常为10个核苷酸），但由于错配，特别是在5′端，它能比一般的十聚体结合更多的位点。其扩增模式被特意设计以利于在长的聚丙烯酰胺凝胶上依片段大小分离时产生一个复杂的梯形带（图7.16）。

与DNA微阵列筛查不同，mRNA差异显示是一种可同时检测多个基因表达的方法，而被跟踪的为之前尚未鉴定的基因。因此，它只是一种表达筛查方法，而不是筛查已知基因表达。它的主要用途为基因的表达比较研究，以明确被比较的两种来源的（例如比较处于不同生理或发育阶段的两类细胞）基因表达将如何改变。这将发现一小部分在不同种类细胞间差异表达的基因。

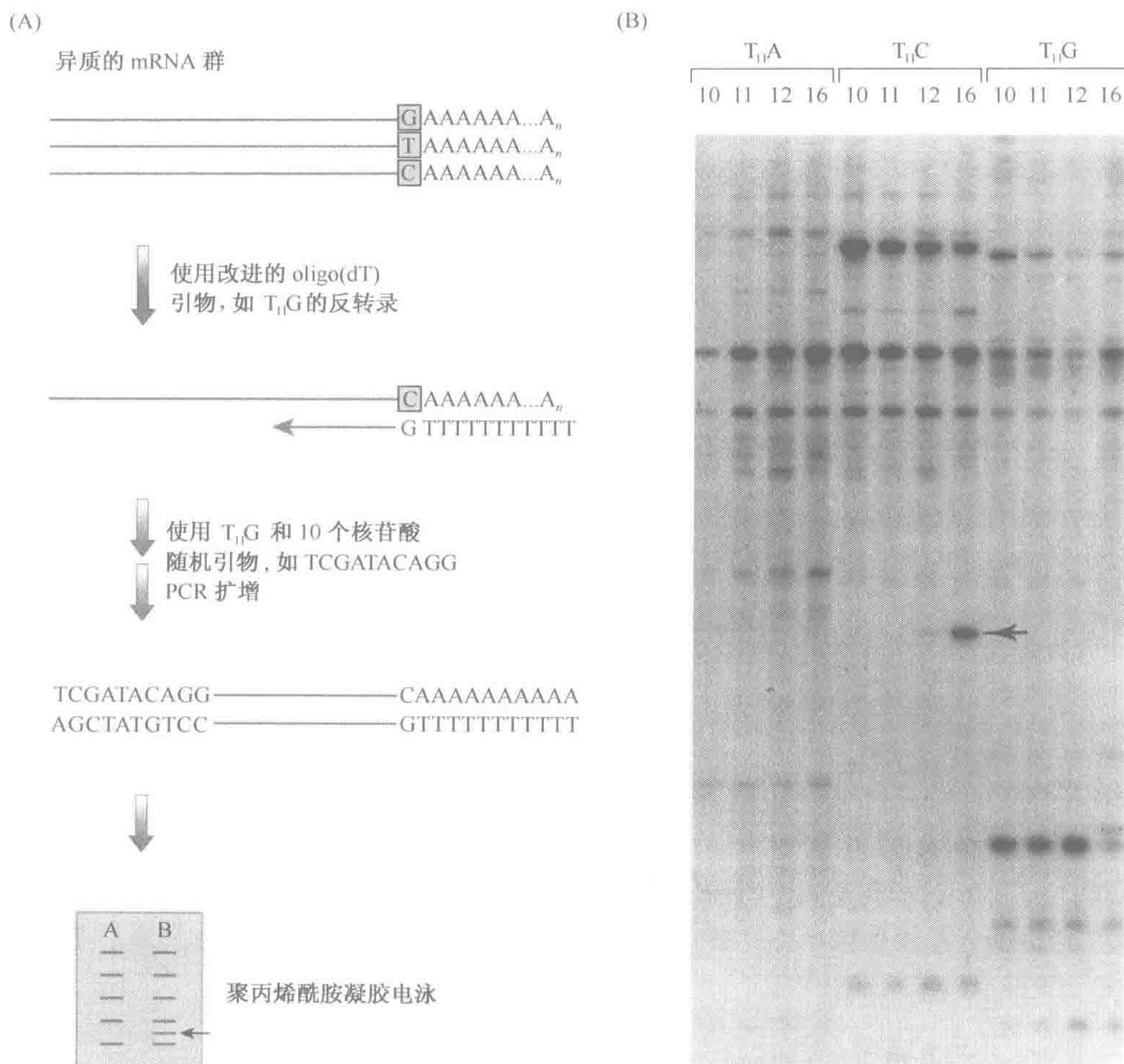


图 7.16　mRNA 差异显示是一种改进的 RT-PCR 方法，用于多个基因的表达筛查

(A) 示意图。来自两种或更多细胞种类的总 RNA 用一个经修饰的 oligo（dT）引物（在本例中为 TTTTTTTTTTTG，或简写为 $T_{11}G$）进行反转录，该引物将引导 cDNA 优先自 poly（A）尾前为 C 的 mRNA 序列合成。扩增反应使用一个随机引物，产物在聚丙烯酰胺凝胶上依片段大小分离。不同 RNA 来源（A 和 B）扩增条带之间的差异提示差异性表达。(B) 一个应用：识别在小鼠心脏不同发育阶段（胚胎第 10，11，12 和 16 天）差异表达的基因。本例采用了三套反应条件，分别使用 T11A，T11C 和 T11G 引物。本图显示了凝胶的一部分，其中可见若干条带在不同发育阶段的强度发生改变。一个变化特别明显的（箭头所示）被证实为 β 珠蛋白。照片由英国 Newcastle-upon-Tyne 大学的 Andy Curtis 和 David Wilson 惠赠。

7.3.4　蛋白质表达筛查通常使用高特异性抗体

因为在蛋白质检测中极具多样性和灵敏性，抗体在研究中有许多用途，而且它们的治疗潜力也相当大（节 21.3.4）。传统上，抗体通过免疫动物分离，但是基因工程制备的抗体被越来越多地使用（框 7.4）。抗体用来以不同的方法检测蛋白质，并且可以使用不同的标记系统。

框 7.4　获取抗体

获取抗体的传统方法

人类基因产物的抗体传统上是通过适当的**免疫原**（immunogen）反复注射合适的动物（如啮齿类、家兔、山羊等）而获得。常使用两类免疫原：

- **合成的多肽**（synthetic peptide）。根据氨基酸序列（由已知的 cDNA 序列推测）设计合成多肽（通常 20～50 个氨基酸）。其思路是，当多肽结合至一个合适的分子（如钥孔血蓝蛋白）时，将形成一种类似于天然多肽片段的构象。这种方法相对简单，但产生合适的特异性抗体的成功率毫无保证且难以预测。
- **融合蛋白**（fusion protein）。另一种方法是将一段合适的 cDNA 序列插入到一个适当的表达克隆载体上经修饰的细菌基因中。其原理是产生一个杂种 mRNA 然后翻译为一种融合蛋白，其 N 端区域来自细菌基因，其余部分来自插入基因（图）。N 端细菌序列通常设计得很短，但仍具有一些优点。例如，它能提供一段信号序列以确保融合蛋白被分泌至胞外基质中，因而使纯化过程简化，还可以保护外源蛋白在细菌内不被降解。因为融合蛋白含有大多数或全部的目的多肽序列，产生特异性抗体的几率可能较高。

这里所示的质粒载体具有一个复制的起点（ori）和一个为了在大肠杆菌中培养而设计的氨苄青霉素耐药基因（amp^R）。多克隆位点（multiple cloning site，MSC）与一个 *LacZ* 基因紧邻，后者可编码 β 半乳糖苷酶，其转录自 *LacZ* 启动子（P_{lac}）按箭头所示方向进行。目的基因（gene X）的一段 cDNA 序列被以适当方向克隆入 MSC 中。经 *LacZ* 启动子表达能产生一种 β 半乳糖苷酶-X 融合体，该蛋白可被大量制备并用来诱导产生蛋白质 X 抗体。一种常用的替代方法是使用 GST **融合蛋白**（fusion protein），其中谷胱甘肽-S-转移酶与目的蛋白偶联，融合蛋白可用亲和色谱法经谷胱甘肽－琼脂糖层析柱容易地纯化。

倘若动物的免疫系统发生应答，特异性抗体将分泌至血清中。收集富含抗体的血清（**抗血清**，antiserum）中包含各种抗体的混合物，每种抗体由一个不同的 B 淋巴细胞产生 [因为免疫球蛋白基因重排是细胞特异性也是细胞类型（B 淋巴细）特异性，节 10.6]。不同抗体将识别免疫原（**多克隆抗血清**，polyclonal antisera）的不同部分（**表位**，epitope）。然而，单一抗体可通过增殖细胞的一个克隆（源于单一 B 淋巴细胞）制备。

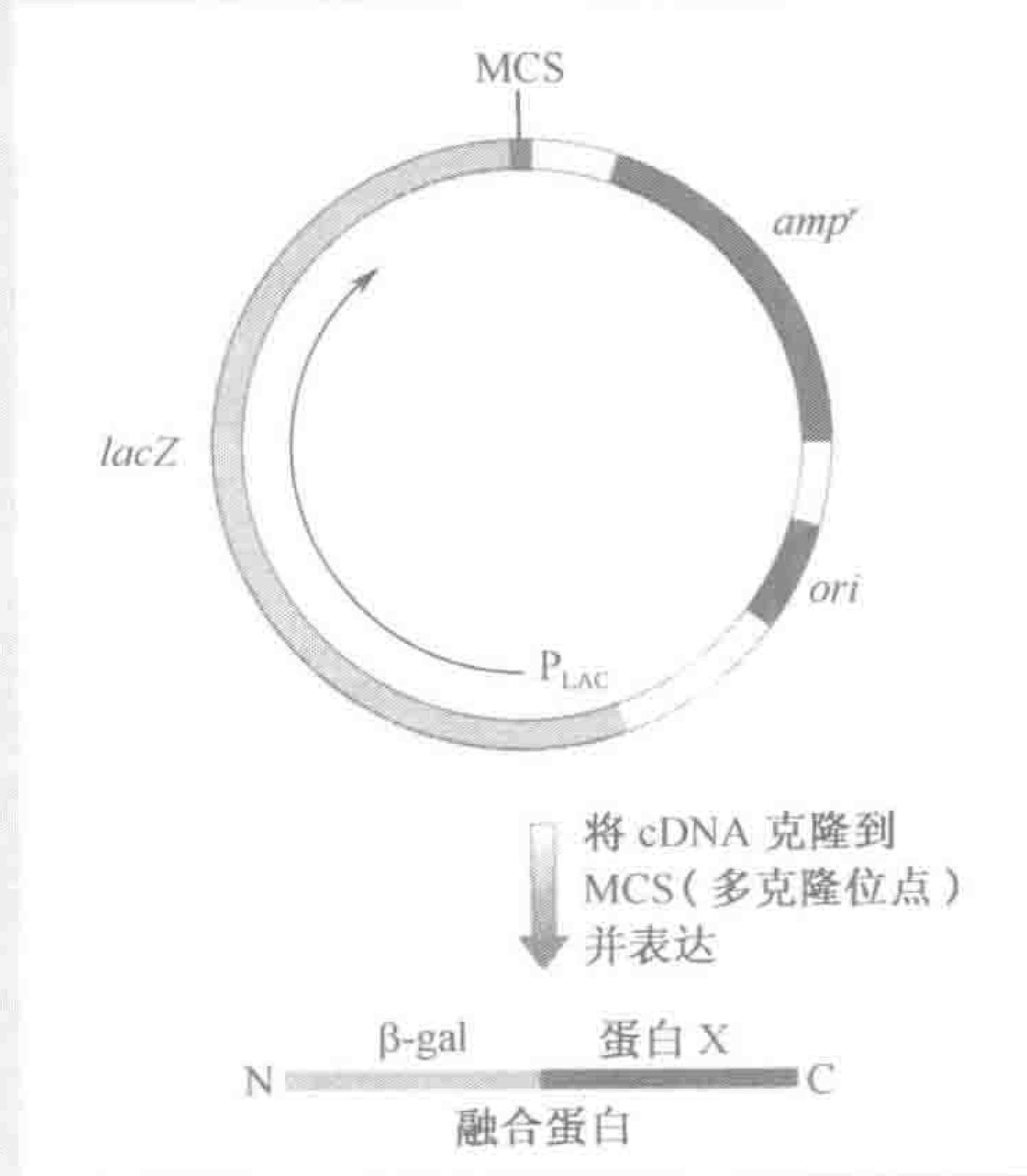
因为 B 细胞在培养中的生存期有限，最好建立永生细胞株：产生抗体的细胞与来自永生化 B 细胞肿瘤的细胞融合。在产生的杂种细胞混合物中选出那些既能产生特定抗体又能在培养中无限增殖的杂种细胞。这样的**杂交瘤**（hybridoma）作为单一克隆繁殖，成为永久而稳定的单一类型**单克隆抗体**（monoclonal antibody，mAb）的来源。

上述制备抗体的方法并不总能保证产生合适的特异性抗体。用于跟踪目的蛋白亚细胞表达的另一种方法是将一个之前获得的抗体结合至一个人工连接的表位（**表位标签**，epitope tagging）来跟踪它。在此过程中，编码已获得抗体的表位序列与编码目的蛋白的序列相连接产生一个 DNA 重组体，其结构与图示大体一致，只是载体系统被设计以适于在拟进行表达研究的哺乳动物或其他细胞中表达。可通过使用表位标签特异性抗体跟踪其蛋白检测重组体在目的细胞中的表达。常用的表位标签如下。

标签序列	来源	位置	mAb
DYKDDDDK	合成的 FLAG	N，C 端	抗 FLAG M1
EQKLISEEDL	人 c-Myc	N，C 端	9E10
MASMTGGQQMG	T7 基因 10	N 端	T7. 标签 Ab
QPELAPEDPED	HSV 蛋白 D	C 端	HSV. 标签 Ab
RPKPQQFFGLM	底物 P	C 端	NC1/34
YPYDVPDYA	流感 HA1	N，C 端	12CA5

框 7.4　获取抗体（续）

Flag，HSV，单纯疱疹病毒。



融合蛋白常设计成免疫原以产生抗体

融合蛋白通常被设计为产生抗体的免疫原。

基因工程抗体

通过上述经典方法制备的抗体均来源于动物。然而，一旦各种免疫球蛋白基因被克隆，DNA 剪切和连接技术可被用于生产新的抗体包括**部分人化抗体**（partially humanized antibody）和**完全人抗体**（fully human antibody，节 21.3.4）。例如，转基因小鼠已被改造为含有人免疫球蛋白基因座，能够在体内产生完全人抗体。

新的方法能完全省去杂交瘤技术和免疫接种的必要。强大的**噬菌体显示技术**（phage display technology）能构建几乎各种针对外来及自身抗原特异性的人类抗体（Winter *et al.*，1994）。该方法的实质在于编码抗体重链和轻链可变序列基因区段被克隆并表达于丝状噬菌体表面，罕见噬菌体可通过与目的抗原结合而从一个混杂的总体中被筛选出来（详细的阐述见节 19.4.6）。

抗体标记和检测系统

抗体可以用不同方法标记，同核苷酸标记和检测一样，抗体可以用于直接或间接检测系统。在**直接检测**（direct detection）方法中，纯化的抗体被报告分子（如荧光素、罗丹明、生物素等；另见节 6.1.2）适当标记后直接与靶蛋白结合。在**间接检测**（indirect detection）系统中，一抗作为一个中介分子并不直接与标记基团相连。

一抗结合到靶点后便与一种同报告基团相连的次级试剂结合。一种常用的次级试剂为 A 蛋白，它是一种在葡萄球菌（*Staphylococcus aureus*）细胞壁中发现的蛋白。由于某些尚不明确的原因，A 蛋白异常牢固地结合于 Ig 重链 Fc 部分第二和第三恒定区域内。一种替代方法是使用**二抗**（secondary antibody，针对主要抗体产生的抗体）。

典型的报告基团可以是一种荧光色素（节 6.1.2）、一种酶（如辣根过氧化物酶、碱性磷酸酶、β 半乳糖苷酶等）或者胶体金。直接检测系统的缺点在于需要将各种各样的一抗与报告分子结合，而间接系统则使用易于获得的经亲和纯化的商品化二抗。用于特殊方法跟踪蛋白表达的标记-检测系统归纳于表 7.1。

表 7.1　用于跟踪蛋白表达的抗体标记-检测方法

标记	检测方法	应用
碘－125	X 光片	免疫印迹
酶	肉眼可检测的产色底物	免疫印迹；免疫细胞化学
生物素	偶联至各种标记的卵白素或链霉亲和素	免疫印迹；免疫细胞化学
荧光染料	荧光显微镜（图 6.5B）	免疫细胞化学；免疫荧光显微镜

免疫印迹（Western 印迹）

这种方法利用依大小分离的细胞提取物检测蛋白总体表达。它常通过一种单向 SDS-PAGE（one-dimensional SDS-PAGE）聚丙烯酰胺凝胶电泳实现，其中蛋白提取混合物先在一种阴离子去污剂——十二烷基硫酸钠（SDS）溶液——中被溶解，该溶液能破坏天然蛋白质中几乎所有的非共价作用。巯基乙醇或二巯基苏糖醇亦被加入以减少二硫键。经电泳分开的蛋白通过适当的染料（例如 Coomassie 蓝）染色或银染观察。

还可以采用双向 PAGE 凝胶电泳：第一个方向为等电聚集，即根据 pH 梯度中的电荷分离，第二个方向与第一个成直角，为利用 SDS-PAGE 依大小分离（Stryer，1995）。这种情况下分开的蛋白转移（"印迹"）到一张尼龙膜上后与特异性抗体结合（图 7.17）。

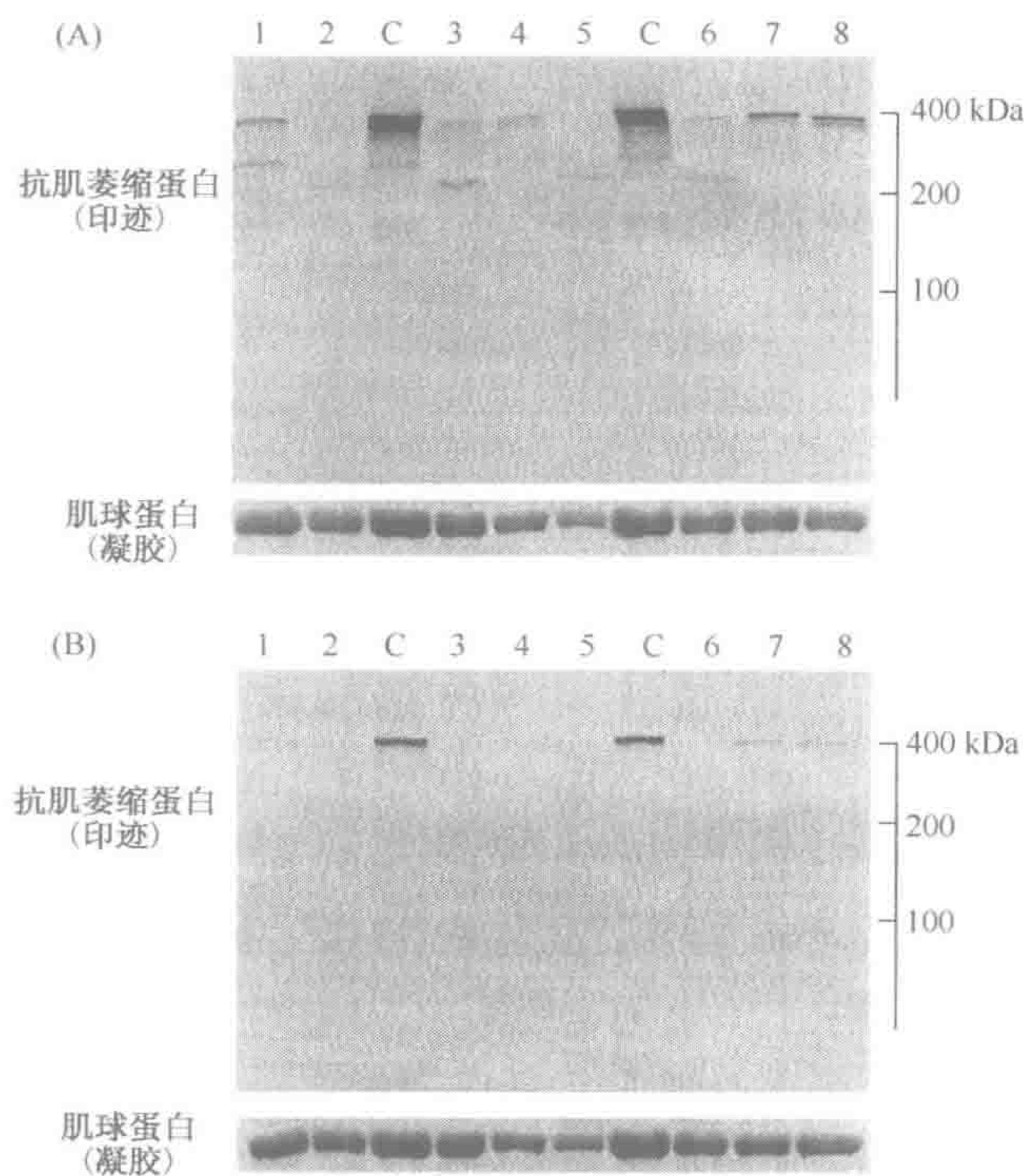


图 7.17　免疫印迹（Western blotting）检测在电泳凝胶中依大小分离的蛋白质

免疫印迹涉及检测在聚丙烯酰胺凝胶中依大小分离并转移（"印迹"）至膜上的多肽。本例显示了该方法在使用两种抗体检测肌营养不良蛋白中的应用。Dy4/6D3 抗体具有杆状结构域特异性，由一种融合蛋白免疫原制备（框 7.4）。Dy4/6D5 抗体具有 C 端区域特异性，由一种合成的多肽免疫原制备。经 BMJ Publishing Group 允许复制于 Nicholson 等（1993）。照片由英国 Newcastle-upon-Tyne 大学的 Louise Anderson（从前的 Nicholson）惠赠。

免疫细胞化学（免疫组织化学）

该技术是在蛋白质水平研究某种组织或多细胞结构的整体表达模式。因而它可以被视为作用于筛查 RNA 表达的组织原位杂交的蛋白质对应方法。组织通常经冰冻或石蜡

包埋后用切片机切成非常薄的切片放到载玻片上。用一种合适的特异性抗体结合组织切片中的蛋白，并产生与邻近组织切片组织学染色有关的表达数据（图 7.18）。

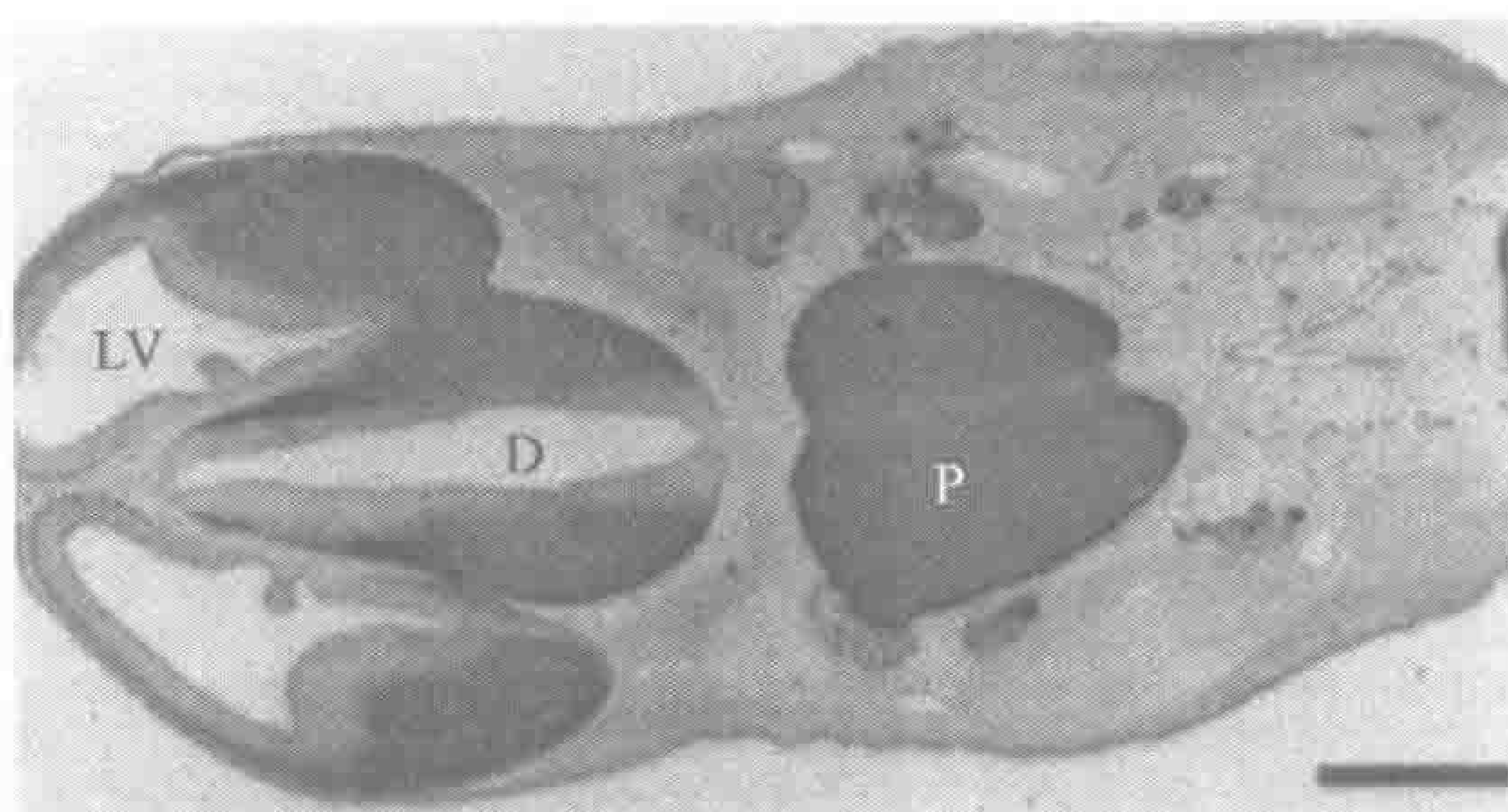


图 7.18 免疫细胞化学

本例为在一个 12.5 天胚胎小鼠大脑横向切片上筛查 β 微管蛋白的表达。抗体检测系统最终用基于辣根过氧化物酶/3，3′二氨联苯胺的褐色反应识别表达。背景组织由 Toluidine 蓝复染显色。缩写：LV，侧脑室；D，间脑；P，脑桥。图片由英国 Newcastle-upon-Tyne 大学 Human Genetics 学院的 Steve Lisgo 惠赠。

免疫荧光显微技术

这种方法被用于研究目的蛋白的亚细胞定位。一种合适的荧光染料如荧光素或罗丹明被偶联至合适的抗体上，使相关蛋白能通过荧光显微技术在细胞内定位（图 6.5B）。

超微结构研究

基因产物或其他分子在细胞内更精细的定位可使用电子显微镜。抗体通常用电子致密的颗粒如胶体性金珠标记。

7.3.5 自体荧光蛋白标签是跟踪蛋白质亚细胞定位的有效手段

绿色荧光蛋白（green fluorescent protein，GFP）是一种最初发现于多管水母（*Aequoria victoria*）中，含有 238 个氨基酸的蛋白质。类似的蛋白表达于许多水母中，并使它们在受到来自荧光素或其他发光蛋白氧化能量的刺激时发出绿光（Tsien，1998）。当 *GFP* 基因被克隆并转染到培养的靶细胞中时，GFP 的异源表达也表现为发出绿色荧光。因此 GFP 是一种**自体荧光蛋白质**（autofluorescent protein）：其自身可作为一种功能性荧光体。由于 GFP 不需要其他试剂，诸如抗体、辅助因子、酶作用底物等，因此可以用作一种独特的报道基因。结果 GFP 可以容易地由常规和共聚焦荧光显微镜跟踪，并成为追踪基因在动物体内表达的一种常用工具（节 20.3.1）。

最成功和广泛的应用是将 GFP 作为融合蛋白中的一个标签偶联到拟跟踪表达的蛋白上。在此情况下主要的目的是研究目标蛋白的亚细胞定位。GFP 本身在细胞内定位并不特异：在绝大多数种类的细胞中 GFP 的荧光均匀地分布于细胞核、细胞质和末端细胞突中。基因工程可制造含有一段 GFP 编码序列的载体，在其中克隆入一段未知蛋白的编码序列——X。产生的 GFP-X 融合体可被转染到合适的靶细胞如培养的哺乳动物细胞中，通过检测 GFP-X 融合蛋白的表达跟踪该蛋白的亚细胞定位。图 7.19 是一个

跟踪由 CLN3 表达产生蛋白质的例子，该基因与一种儿童神经退行性病变 Batten 病相关。

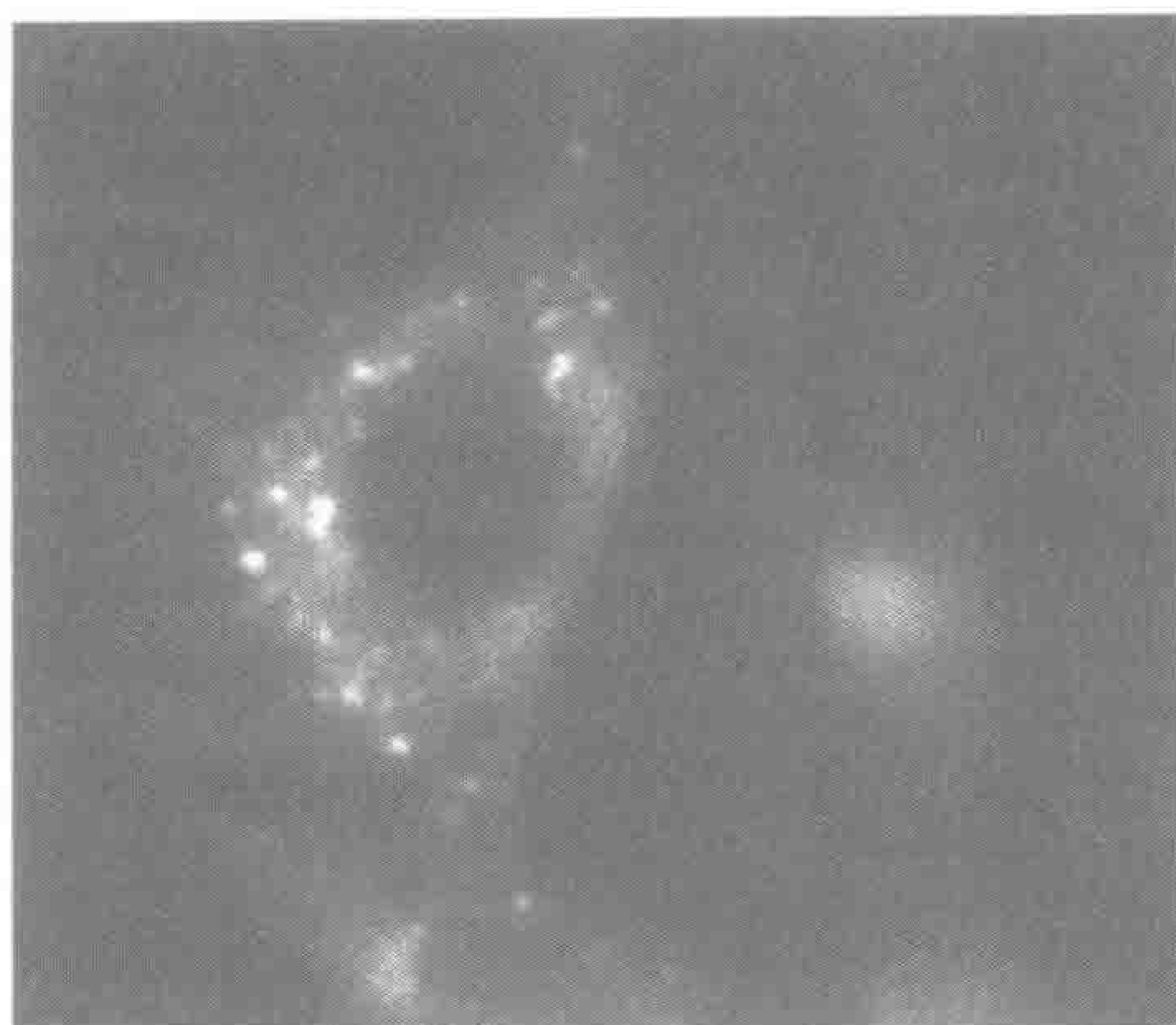


图 7.19　以绿色荧光蛋白作为标签是一种跟踪蛋白表达的有效手段

本例显示了使用 GFP 标签的 Batten 病蛋白在瞬时转染的活 HeLa 细胞中的表达。CLN3 被克隆入 GFP 表达载体 pEGFP-N1，因而产生了一个由 Batten 病蛋白和偶联至其 C 端的 GFP 序列（一个 GFP 标签）构成的融合蛋白。本例显示了部分 HeLa 细胞中 CLN3p/GFP 以囊泡点式分布于整个细胞质。该实验及其他分析结果表明 Batten 病蛋白是一种 Golgi 整合膜蛋白。经 Oxford University Press 允许，复制于 Kremmidiotis 等（1999）。Hum. Mol. Genet. 8，523-531。

（刘洪　译）

进一步阅读

Sambrook J, Russell D (2001) *Molecular Cloning: A Laboratory Manual*, 3rd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

参考文献

Chee M, Yang R, Hubbell E *et al.* (1996) Accessing genetic information with high-density arrays. *Science* **274**, 610–614.

Church DM, Stotler CJ, Rutter JL, Murrell JR, Trofatter JA, Buckler AJ (1994) Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* **6**, 98–105.

Clement-Jones M, Schiller S, Rao E *et al.* (2000) The short stature homeobox gene *SHOX* is involved in skeletal abnormalities in Turner Syndrome. *Hum. Molec. Genet.* **9**, 695–702.

Daniels R, Kinis T, Serhal P, Monk M (1995) Expression of myotonin protein kinase gene in preimplantation human embryos. *Hum. Molec. Genet.* **4**, 389–393.

Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of single RNA transcripts *in situ*. *Science* **280**, 585–590.

Frohman MA, Dush MK, Martin GR (1988) Rapid production of full length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85**, 8998–9002.

Ginsburg M (1994) In: *Guide to Human Genome Computing* (ed. MJ Bishop), pp. 215–248. Academic Press, New York.

Hauge Y, Litt M (1993) A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. *Hum. Mol. Genet.* **2**, 411–415.

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919.

Kremmidiotis G, Lensink IL, Bilton RL, Woollatt E, Chataway TK, Sutherland GR, Callen DF (1999) The Batten disease gene product (CLN3p) is a Golgi integral membrane protein. *Hum. Mol. Genet.* **8**, 523–531.

Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single cell gene expression profiling. *Science* **297**, 836–840.

Liang P, Averboukh L, Pardee AB (1993) Distribution and cloning of eukaryotic mRNAs by means of differential display: refinements and optimization. *Nucleic Acids Res.* **21**, 3269–3275.

Lovett M (1994) Fishing for complements: finding genes by direct selection. *Trends Genet.* **10**, 352–357.

Maxam AM, Gilbert W (1980) Sequencing end labeled DNA with base-specific chemical cleavages. *Methods Enzymol.* **65**, 499–560.

Meldrum D (2000) Automation for genomics, part two:

sequencers, microarrays and future trends. *Genome Res.* **10**, 1288–1303.

Monaco AP (1994) Isolation of genes from cloned DNA. *Curr. Opin. Genet. Dev.* **4**, 360–365.

Needleman SB, Wunsch CD (1970) A general method applicable to the search of similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443–453.

Pykett MJ, Murphy M, Harnish PR, George DL (1994) The neurofibromatosis 2 (NF2) tumor suppressor gene encodes multiple alternatively spliced transcripts. *Hum. Mol. Genet.* **3**, 559–564.

Riordan JR, Rommens JM, Kerem B *et al.* (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–1073.

Schutze K, Lahr G (1998) Identification of expressed genes by laser-manipulated manipulation of single cells. *Nature Biotechnol.* **16**, 737–742.

Simone NL, Bnonner RF, Gillespie JW, Emmert-Buck MR, Liotta LA (1998) Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **16**, 272–276.

Stryer L (1995) *Biochemistry*, 4th Edn. W.H. Freeman & Co., New York.

Tonkin E, Hagan DM, Li W, Strachan T (2002) Identification and characterisation of novel mammalian homologs of *Drosophila* polyhomeotic permits new insights into relationships between members of the polyhomeotic family. *Hum. Genet.* **111**, 435–442.

Tsien RY (1998) Green fluorescent protein. *Ann. Rev. Biochem.* **67**, 509–554.

Waterman MS, Smith TF, Beyer WA (1976) Some biological sequence metrics. *Adv. Math.* **20**, 367–387.

Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR (1994) Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433–455.

第8章　基因组计划和模式生物

本章内容

8.1　基因组计划的开创性意义

8.1.1　基因组计划为宇宙内部的系统研究提供了手段

经过许多世纪的探索，我们对外部宇宙中至少比较易于接近的部分有了大致的了解。这个领域给我们留下了深刻的印象，一些观念必然超出了我们正常的阅历（诸如现在被认为存在的 13 维!）。但是，在我们内部也有一个大的、尚未探索的宇宙，它同样是一个令人惊叹的领域（人脑的复杂性就是一个有用的例子：大约 10^{11} 个神经元以及在 10^{15} 个互相联系的区域中的某个部位）。

直至现在，对我们内部宇宙的探索性旅程在规模上一直是适度的和局限的。运用显微镜研究细胞及亚细胞结构提供了进入这个内部世界的一条主要途径，生物化学和分子、细胞生物学的开拓性进展紧随其后。新世纪伊始，我们可以从这些初始的研究成果出发，平稳地过渡到一个整体上更高的阶段。现在，对我们的内部宇宙进行严谨而系统的探索将成为可能。

为这个新的探索阶段铺平道路的催化剂是**人类基因组计划**（Human Genome Project，HGP）——一项真正的国际性项目。它正式启动于 1990 年，是生物学的第一个“大项目”，计划用时 15 年。HGP 和其他基因组计划率先寻求了解详细说明生物体的精

细化学指令，以及全部基因组序列（基因组是指整套不同的 DNA 分子——见框 8.1 基因组学专业词汇表）。对于许多科学家来说，这是（元素）周期表的生物学等价物：所有的物质都可以归结到一个元素周期表中，但在一更高的水平上，每一个生物体可以归结到基因的一个周期表中。

框 8.1　基因组学词汇表

厘摩（centiMorgan，cM）。**遗传图**（genetic map）中的距离单位（见下文）。在人类基因组中，1cM 大约相当于物理图上 1Mb 的距离。

厘伦琴（centiRay，cR）。**辐射杂种细胞图**（radiation hybrid map）的图距单位（见下文）。

克隆（clone）。DNA 克隆是用细胞克隆方法（节 5.3.1）或 PCR（节 5.2.1）纯化得到的相同 DNA 分子群。

叠连群（contig）。表现为含有来自某染色体相邻及重叠区的插入 DNA 分子的一系列 DNA 克隆（框 8.5）。

DNA 标记（DNA marker）。位于或可能位于遗传图［对于**多态性标记**（polymorphic marker）来说——见下文］或物理图（对于全部标记来说）的一段 DNA 序列的通称。

CpG 岛（CpG island）。富含 GC 的 DNA 短片段，一般＜1kb，含有许多非甲基化的 CpG 二核苷酸。CpG 岛倾向于标示基因 5′端（框 9.3）。

DNA 文库（DNA library）。DNA 克隆的集合，意即全面地代表一起始的 DNA 群，对于**基因组 DNA 文库**（genomic DNA library）来说，起始 DNA 是来自一既定细胞群体（它表示在不同细胞类型之间很少有变异）的 DNA 总和。至于 **cDNA 文库**（cDNA library），起始 DNA 是用反转录酶由一特定组织（不同组织的 cDNA 有相当大的变异）的单链 RNA 制备的。如何制备和筛选文库见节 5.3.4 和节 5.3.5。

表达序列标签（expressed sequence tag，EST）。随机选择一 cDNA 克隆进行测序并设计特异性 PCR 扩增相应基因组 DNA 片段的特异性引物而得到的一个表达的**序列标签位点**（sequence tagged site，STS，见下文）。

遗传图（genetic map）。依靠追踪世代中表型和（或）多态性标记的遗传而绘制的图。彼此相关的多态性位点的定位是根据它们在减数分裂过程中重组的频率。距离单位为 1 厘摩（1cM），表示重组率为 1%。

基因组（genome）。特定种族细胞中各种 DNA 分子的总称。在人类，基因组由 25 种不同的 DNA 分子组成：单一类型的线粒体 DNA 和 24 种不同的核 DNA 分子（节 9.1.1）。但由于核 DNA 的数量很大，所以基因组这个词汇也常用来泛指核 DNA 分子的集合［更准确应称为**核基因组**（nuclear genome）；线粒体 DNA 常称为**线粒体基因组**（mitochondrial genome）］。

杂种细胞作图（hybrid cell mapping）。用各种杂种细胞嵌板可以把人类 DNA 标记指定到特定的染色体或染色体亚区位置，该嵌板含有整套的啮齿类染色体和各种人类染色体亚群或 X 线碎裂的人类染色体片段［**辐射杂种细胞**（radiation hybrids），框 8.4］。

微卫星标记（microsatellite marker）。一类常用的 DNA 标记，主要是因为这种标记是高度多态性的（图 7.7 和图 7.8）。

物理图（physical map）。一提供 DNA 分子线性结构信息的图。最精细的物理图是核苷酸序列。

多态性标记（polymorphic marker）。多态性的（遗传）标记是显示个体间变异的 DNA 序列，并可用来通过追踪大家系中等位基因如何分离来构建遗传图。标记定位于编码序列或其他基因组分，但大多数位于非编码 DNA 中。虽然过去使用 RFLP 甚至蛋白质多态性，但常用的标记是微卫星和 SNP。

框 8.1　基因组学词汇表（续）

辐射杂种细胞图 [radiation hybrid (RH) map]。彼此相关的 STS 依据辐射诱导染色体断裂引起 STS 分离的频率来定位的一种基因组图。通过分析**杂种细胞系**（hybrid cell）（人类一仓鼠）嵌板来测定频率，杂种细胞系含有最初暴露于 X 射线产生的不同样式人类染色体片段。图距单位是 1 厘伦琴（1cR），表示两个位点间发生断裂的概率是 1%。

限制性片段长度多态性（restriction fragment length polymorphism，RFLP）。一类过去广为应用，但现在因为它们通常多态性低并且不易分型而现在很少使用的 DNA 标记（节 7.1.3）。

单核苷酸多态性（single nucleotide polymorphism，SNP）。SNP 提供了一类正被日益应用的 DNA 标记。它们在 DNA 中频繁出现并且容易采用自动化方法分型，可以一次分析非常大量的样本。SNP 如何分型见节 7.1.3 和框 18.2。

序列标签位点（sequence tagged site，STS）。独一无二的存在于某个基因组内的任一短（一般<500bp）序列，已设计引物能够特异性 PCR 扩增此序列（框 5.4）。STS 常由随机测序基因组克隆的末端而设计，故通常是非多态性的，但已知一亚组 STS 是多态性的，包括**微卫星标记**（microsatellite marker）（见上文）。

人类基因组计划（HGP）的目标

HGP 的主要原理是获得关于我们的遗传组成的基本信息，它将增进我们对于人类遗传学以及各种基因在健康和疾病中作用的科学认识。自 1981 年人类线粒体 DNA（mtDNA，16.5kb）序列公布以来（节 9.1.2），HGP 的主要目标就是对非常大的核基因组（大约 3000Mb）进行测序。要实现这一目标，首先需要获得高分辨率的人类遗传图，然后以此为支架（框架）构建高分辨率物理图，最后获得最终的物理图，即人类基因组的完整序列。

除了对人类核基因组进行作图和测序这一主要目标之外，HGP 在其开始时设想了一系列辅助计划：

- **适当技术和工具的开发**。这包括开发遗传和物理作图方法、DNA 测序技术、数据库的设计和建立、序列分析信息学等等。
- **五种模式生物的基因组计划**：大肠杆菌（*E. coli*），酿酒酵母（*Saccharomyces cerevisiae*），秀丽新小杆线虫（*Caenorhabditis elegans*），果蝇（*Drosophila melanogaster*）和小鼠。在这里，目的是双重的：提供关于这些模式生物的更多的必需信息；为执行和改进 HGP 所需的各种技术、工具提供试验病例。
- **伦理、法律和社会意义**。全部经费的重大部分已投入到这一重要而又容易被忽略的领域。

到 2003 年，人类基因组和五种最初的模式生物基因组的测序目标已经实现，而且序列可以从互联网上获得。HGP 期间，许多其他模式生物的基因组计划也已经启动，到 2003 年 5 月已确定了 140 种其他生物的基因组序列（节 8.4）。额外的辅助研究一直关注人类基因组内部序列变异的范围。

8.1.2　期望基因组计划的医学及科学利益是巨大的

对许多人类生物学家和遗传学家来说，HGP 是一项令人激动的历史性使命。一个

主要的理由是预期的医学利益（Collins and Mckusick，2001；Suburamanian *et al*.，2001；van Ommen，2002）。对于一个主要的单个基因引起的遗传病来说，对判定为具有携带一个致病基因风险的个体进行全面的产前/症状前疾病诊断将成为可能。另外，基因结构的信息将用于研究各个基因如何起作用以及如何调节，这些信息将为人类生物过程提供非常需要的解释。它还将为开展新的疾病治疗策略提供准则，并拓展现有的治疗方案。突变筛查的广泛应用也将为医疗保健方法带来根本性的改变，它根据个体风险的鉴别，更多的从治疗进展期疾病的方法转变为预防疾病的方法（**个体化医学**，personalized medicine）。

然而尽管这些可能性是令人激动的，但在精确地、全面地了解一些基因如何发挥功能以及如何调节方面还有许多意想不到的困难（前车之鉴就是在获得相关序列几十年后，在由氨基酸序列预测蛋白质结构的缓慢进展，以及缺乏理解珠蛋白基因表达调节的精确方式是如何协调的）。另外，应该是开展新的治疗最容易的靶标的单基因病是很罕见的；大多数常见疾病是多因素的，而且面临着相当大的挑战。因此，虽然人类基因组计划收集的数据必然很有医学价值，但要实现一些重要的医学应用还需要相当长的时间。

当我们步入**后基因组时代**（post-genome era）时，国际间的重大协作正集中于人类基因组序列如何能够详细说明一个个体，其他有机体的 DNA 与我们以及其生物学如何相关。通过研究基因组结构（**常规的基因组学**，conventional genomics）获得的序列信息为其他研究基因组功能（**功能基因组学**，functional genomics）以及不同基因组如何彼此相关（**比较基因组学**，comparative genomics）的大规模方法铺平了道路。这些主题包括在 19 章和节 12.3。

8.2　人类基因组计划的研究背景和组织机构

8.2.1　DNA 多态性和新的 DNA 克隆技术为基因组测序铺平了道路

自 20 世纪 50 年代中期以来，我们已经获得了一个人类基因组的粗略物理图，即根据大小和形态来区分染色体的细胞遗传图。为了发展更为精细的物理图需要更新的方法。主要的问题在于获得人类遗传图并以此为支架建立更为精细的物理图的最初目标似乎是不可能的。反而，当几十年前构建了果蝇和小鼠的**经典遗传图**（classical genetic map），随后不断完善时，人类遗传学家非常羡慕地关注它。

经典遗传图谱是建立在**基因**（gene）的基础上。它是通过杂交不同的突变体从而判断两个基因座是否关联而构建的。然而，一个经典的人类遗传图是绝不可能获得的，因为具有不同遗传病的两个个体婚配的概率非常小。没有一个遗传图提供锚定点，很难想像如何获得全部染色体精细的物理图。

在 20 世纪 70 年代晚期，一个转折点是认知不断增加——事实上，是绝大多数人类基因组序列的变异发生在基因外部并且能被检测。直到那时一直备受关注的变异仍集中于蛋白质多态性。由于编码 DNA 只占基因组的一个非常小的部分（2%），并且不易发生变异（因为它在功能上很重要并且在进化过程中高度保守），所以只能研究少数几种蛋白质标记。相反，绝大多数超过 98%的非编码 DNA 保守性不是很好，而且 DNA 序

列易于发生变异。

20 世纪 70 年代晚期，首次拥有了分析 DNA 变异的方法（通过筛查**限制性片段长度多态性**，resriction fragment length polymorphism，RFLP，框 8.1）。最终，构建一个全面的、非经典的人类遗传图的想法成为可能（Botstein *et al.*,1980）。从现在开始，人类遗传学家可以利用随机散布于我们基因组中的 DNA 标记来绘制日益详尽的遗传连锁图。在家系研究中，通过检验并观察来自两个或更多的标记的特异性等位基因是否共同分离，就可以把 DNA 标记定位于一特定的**连锁群**（linkage group）。通过一个或更多的组成性标记的物理作图，就可以把各个连锁群依次定位到特定的染色体上［例如通过标记某个标记并将其与分裂中期染色体杂交（图 2.16 和图 2.17）或使用杂种细胞嵌板（节 8.3.2）］。

另一个重要的必要条件就是发展强大的 DNA 克隆技术，以获得含有大片段 DNA 的克隆群（**大插入子 DNA 文库**，large insert DNA library）。克隆的插入片段已经通过随机切割全基因组获得，但可以检测并观察它们是否含有哪个特殊的 DNA 标记以及是否与其他克隆的插入片段共享标记。如果它们确实共享，那么这些克隆通常含有重叠的插入 DNA 序列，就有可能根据 DNA 插入子间的重叠部分，即所谓的**克隆叠连群**（clone contig）来制作 DNA 克隆的有序图谱。

8.2.2　人类基因组计划主要在具有高通量测序能力的大型基因组中心进行

人类基因组计划的组织机构

当美国人类基因组计划提供了由政府资助人类基因组计划的初始契机时，几个其他国家迅速开展了他们自己的人类基因组计划。英联邦和法国基因组中心迅速取得了成绩，最近一些其他国家的基因组中心也做出了相当大的贡献，特别是日本和德国。为了协调不同国家的努力，1988 年成立了**人类基因组组织**（Human Genome Organization，HUGO），呈交当局解决的事情包括促进资源交换，鼓励公开辩论，以及对人类基因组研究的意义提供建议（Mckusick，1989）。



图 8.1　Wellcome Trust Sanger 研究所中的大规模 DNA 测序
位于英联邦 Hinxton 的 Wellcome Trust Sanger 研究所是由政府出资的人类基因组计划的一个最大贡献者。其资料可在http：//www.sanger.ac.uk 得到。

由于研究涉及领域较广，所以许多人类基因组测序技术都集中在少数几个非常大的具有工业化测序能力的**基因组中心**（genome center）（图 8.1）。自动荧光标记 DNA 测序成为常规方法，而**毛细管 DNA 测序**（capillary-based DNA sequencing）（节 7.1.2）的出现，为高通量测序提供了更多必需的帮助。对于由政府出资的 HGP 来说，大部分序列是由 5 大中心贡献的，英联邦的 Wellcome Trust Sanger 研究所和 4 个美国的实验中心：马萨诸塞的 Whitehead 研究所/马萨诸塞技术研究所，华盛顿大学，DoE Joint 基因组研究所和 Baylor 医学院。与这些和一些其他大型基因组中心合作的是由小型实验室构成的世界范围的网络体系，这些小实验室大多数致力于定位和鉴别致病基因并代表性地集中于非常特异的染色体亚区（图 8.2）。

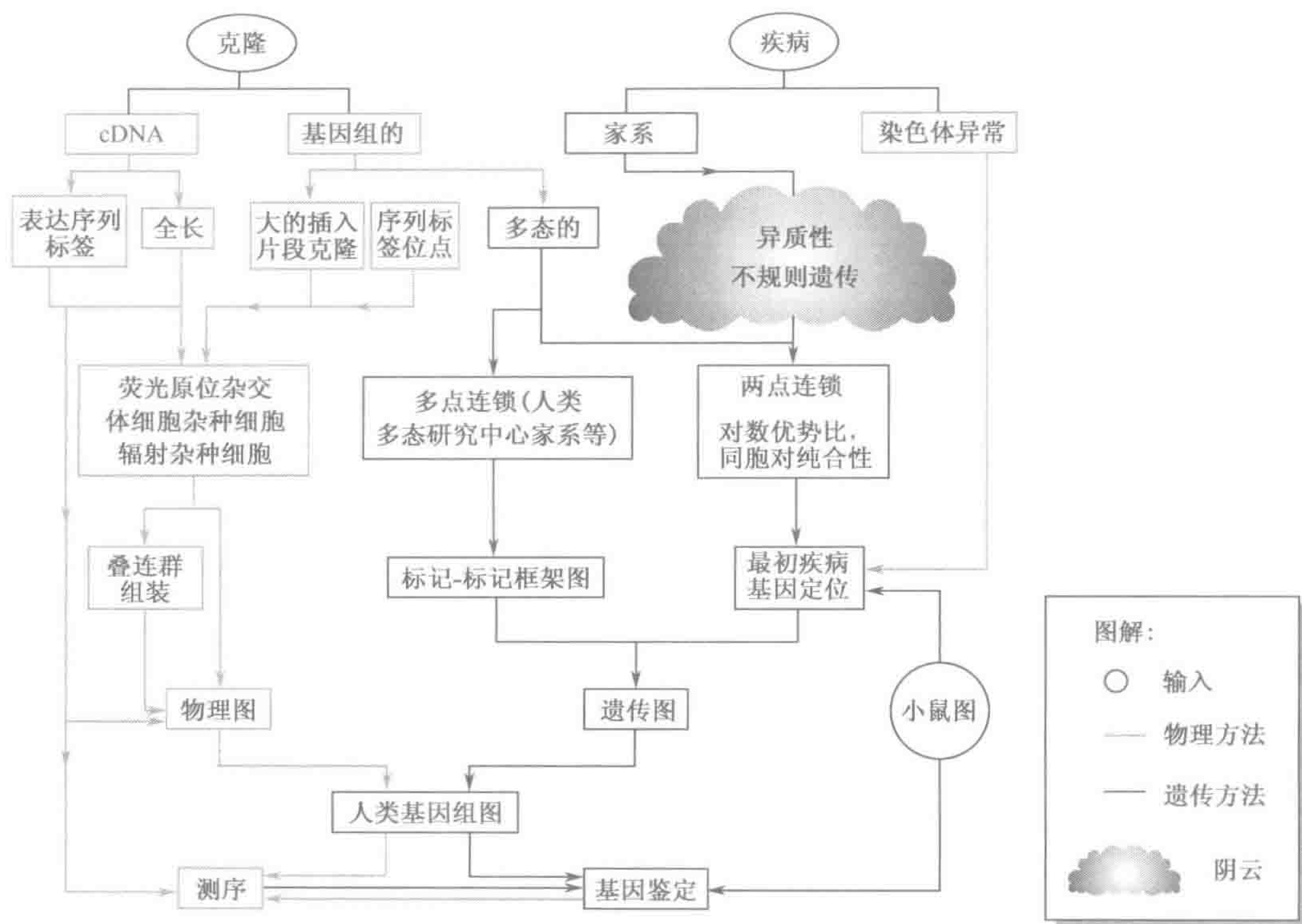


图 8.2 人类基因组计划中使用的主要科学策略和方法

人类基因组计划主要的科学动力始于人类基因组和 cDNA 克隆的分离（通过细胞克隆或 PCR 克隆）。随后这些技术用于构建高分辨率的遗传图和物理图，并以此获得最终的物理图—3000Mb 的核基因组的完整核苷酸序列。该计划必然会影响到定位和鉴别人类疾病基因的研究。另外，辅助性计划包括研究遗传变异（人类基因组多样性计划）（节 8.3.7），模式生物基因组计划（节 8.4）以及伦理、法律和社会意义的研究。获得的数据用于作图，并输入序列数据库用来进行快速电子查询和数据分析。EST，表达序列标签；STS，序列标记位点。

在基因组中心网络与相互合作的实验室之间的交流极大程度地依赖电子交流，将来也会一直如此。处理和储存大量快速产生的测序数据的需要导致了大型的**电子数据库**（electronic database）的发展，其中至少由政府出资的作图和测序结果是可以在互联网上免费获得的。随后可以在世界各处的远程计算机终端进行数据分析。根据输入数据的

来源，有两种相关的数据库：

- 储存全球作图和测序数据的中心储存处。全部 DNA 和蛋白质**序列数据库**（sequence database）在基因组计划启动的几十年前就建立了，但专门的种属特异性作图数据库则是最近才建立的，如**基因组数据库**（genome database，GDB），一个专门用于存储人类作图数据的数据库（**注**：不同物种的 DNA 片段和基因的命名各有其特殊的专有名词—人类专有名词见进一步阅读及框 8.2）。
- 储存本地产生数据的数据库。为了提高他们自己的工作效率，大型基因组作图和测序中心把他们自己实验室完成的数据储存到专用的数据库。与数据输入不同，这些数据可以从由政府出资的基因组中心的网络系统免费获得。

框 8.2　人类基因和 DNA 片段专有名词

使用的专有名词是由 HUGO 命名委员会规定的。基因和假基因通常用 2～6 个字符组成的符号表示：假基因序列用 P 及相应基因符号表示。对匿名的 DNA 序列，通常用 D（=DNA）加 1～22、X 或 Y（表示染色体位置），然后是 S 表示一独特的片段，Z 表示一染色体特异性重复 DNA 家族或者 F 表示多位点 DNA 家族，最后是序列号。匿名的 DNA 序列的编号后面加字母 E 表示已知该序列有表达。

符号	说明
CRYB1	晶体蛋白 β 多肽 1 基因
GAPD	甘油醛-3-磷酸脱氢酶基因
GAPDL7	GAPD 样基因 7，功能状态不明
GAPDP1	GAPD 假基因 1
AK1	腺苷酸激酶基因，位点 1
AK2	腺苷酸激酶基因，位点 2
*PGK1 * 2*	*PGK1* 基因座的第二等位基因
B3P42	3 号染色体上的 42 号断点
DYS29	独特的 DNA 片段，编号 29，位于 Y 染色体
D3S2550E	独特的 DNA 片段，编号 2550，位于 3 号染色体，已知有表达
D11Z3	11 号染色体特异性重复 DNA 家族，编号 3
DXYS6X	X 染色体上发现的 DNA 片段，Y 染色体上有已知的同源序列，是被分类的第 6 对 XY 同源对
DXYS44Y	Y 染色体上发现的 DNA 片段，X 染色体上有已知的同源序列，是第 44 对 XY 同源对
D12F3S1	12 号染色体上的 DNA 片段，多位点家族 3 的第一号成员
DXF3S2	X 染色体上的 DNA 片段，多位点家族 3 的第二号成员
FRA16A	16 号染色体上的脆性位点 A

8.3　人类基因组是如何作图及测序的

官方的人类基因组计划预计从1990～2005用时15年，但实际进程比预想的要快。遗传图的进程比原计划提前，最后阶段的大规模DNA测序因为自动荧光DNA测序技术的发展以及与私人公司Celera的竞争所产生的动力而得以推进。至2003年，所有人可以从互联网上获得基本完成的序列。一些主要里程碑的时间表见框8.3。

框8.3　人类基因组作图和测序中的主要里程碑

1956：第一个人类基因组的物理图被确定——染色组织的光学显微镜学显示我们的细胞含有46条染色体，共有24种不同类型的染色体。

1977：英联邦剑桥大学的Fred Sanger及其同事们公布了双脱氧DNA测序方法，这个方法在超过四分之一世纪之后仍是现代DNA测序的基础。

1980：Botstein等（1980）提出使用一组随机的DNA标记（RFLP），可以构建人类遗传图。

1981：Fred Sanger及其同事们公布了人类线粒体DNA的完整序列（节9.1.2）。

1984：美国能源部（DoE）部分资助的犹他州Alta工作室，来评价突变检测和描述极性的方法以及规划未来技术。一个重要结论就是，高效的突变检测需要一个庞大、复杂、花费昂贵的测序计划。

1987：关于人类基因组启动计划的美国能源报告部设想了三个主要目标：人类染色体精细物理图的制作；支持人类基因组研究的技术和设备的发展；以及通信网络与运算能力和数据库容量的扩展。

1988：美国国立卫生研究院（National Institutes of Health，NIH）成立人类基因组研究办公室（以后更名为国家人类基因组研究中心National Center for Human Genome Research）来协调NIH与其他美国机构在基因组活动中的合作。同年为了协同国际间工作，人类基因组组织（Human Genome Organization，HUGO）成立，呈交当局解决的事情包括推动研究资源互换，鼓励公开辩论以及对人类基因组研究意义提供建议（McKusick，1989）。

1990：**人类基因组计划**（Human Genome Project，HGP）在美国正式启动，将耗资30亿美元，历时15年。

1991：基因组数据库（Genome Database，GDB）——人类DNA绘图数据储存库成立。

1992：法国Généthon实验室的Jean Weissenbach及其同事们公布了第一个基于微卫星标记制作的全面的人类遗传连锁图（Weissenbach *et al.*，1992）。

1993：法国Généthon实验室的Daniel Cohen及其同事们公布了第一代基于大插入片段DNA克隆制作的物理图（Cohen *et al.*，1993）。

1995：Whithead学院/马萨诸塞科技学院的Eric Lander及其同事们公布了第一个基于序列标签位点制作的详细的物理图（Hudson *et al.*，1995）。

1998：英联邦Sanger中心领导的国际协作组公布了基因图98，它是第一个相当全面的基因标记图（Deloukas *et al.*，1998）。

1999：英联邦Sanger中心领导的国际协作组公布了第一个人类染色体——22号染色体——的完整序列（Dunham *et al.*，1999）。

2001：由政府出资的国际研究协作组和私人公司Celera公布了人类基因组序列的粗略工作草图（约含全部序列的90%）（International Human Gene Sequencing Consortium，2001，Venter *et al.*，2001）。

2003：人类基因组测序完成。

8.3.1　第一个可用的遗传图是基于微卫星标记制作的

20 世纪 80 年代早期，人们就意识到现在可以获得一张全面的人类遗传图，这种意识鼓舞人们努力地去构建它。1987 年首张这样的图问世，其主要是基于**限制性片段长度多态性**（restriction fragment length polymorphism，RFLP）绘制的。尽管这个成果很伟大，但 RFLP 图具有严重的局限性：标记之间的平均间距相当大，更主要的是 RFLP 标记信息量不足（只有两个等位基因），且不易分型。

因此，人们的注意力转移到用微卫星标记作图上，这些标记具有信息量高，易分型，散布于全基因组的优点（节 9.4.3 和图 7.7、图 7.8）。随后的 5 年内，法国 Généthon 实验室的研究者们报道了首张基于微卫星的人类基因组连锁图（Weissenbach *et al.*,1992）；两年后一个国际协作组公布了一个进一步改良的图，该图主要依据高密度微卫星标记，大约每厘摩（cM）含有一个标记（Murray *et al.*,1994）。

1994 年公布的高分辨人类遗传图满足了人类基因组计划的第一个主要科学目标，并在现在为发展全部染色体的精细物理图提供了一个重要框架。从现在开始，HGP 的主要焦点将是发展和完善物理图以得到最终的物理图，即每条染色体的全部 DNA 序列（节 8.3.2）。然而人类基因组的遗传作图仍以两种主要方式进行：

- 微卫星图的完善。目前最精细的图是由冰岛的 deCODE 遗传学研究组构建的。它涉及 146 个家系中 5136 个微卫星标记的分型，总计 1257 次减数分裂事件（Kong *et al.*,2002）。
- 单核苷酸多态性图的发展（节 8.3.7）

8.3.2　人类基因组首张高分辨率物理图是基于克隆叠连群和 STS 界标制作的

人类基因组物理绘图的一般方法

虽然构建不同的人类遗传图使用不同类型的标记，但都有一个共同的潜在原则——标记在各种多代家系成员中分型，以及数据输入计算机以检查与等位基因共同分离的标记。由于可能有多种不同类型的图，所以物理作图也不同（表 8.1）。人类基因组首张物理图是基于染色体显带技术绘制的并且在 40 多年前就绘制成功（染色体显带图的现代实例，图 2.14）。

表 8.1　不同种类的物理图可用于人类核基因组作图

图的类型	实例/方法学	分辨率
细胞遗传图	染色体显带图	每带平均为几个 Mb 的 DNA
染色体断裂点图	体细胞杂种细胞嵌板含有来自自然易位或缺失染色体的人类染色体片段	一条染色体上相邻染色体断裂点之间的距离通常是几个 Mb
	单染色体辐射杂种细胞（RH）图	断裂点间距通常是很多 Mb
	全基因组 RH 图	分辨率高达 0.5Mb
限制性酶切图	**罕见-切割**（rare-cutter）限制性酶切图，如 *NotI* 图	罕见—切割限制性酶切图为几百个 kb

续表

图的类型	实例/方法学	分辨率
克隆叠连群图	重叠 YAC 克隆	平均 YAC 插入子为几百个 kb 的 DNA
	重叠黏粒克隆	平均黏粒插入子为 40kb
STS（序列标签位点）图	需要有序克隆的预先序列信息来对 STS 进行排序	可能小于 1kb，但标准 STS 图分辨率为几十 kb
EST（表达序列标签）图	需要测序 cDNA 然后再把 cDNA 定位到其他物理图中	人类核基因组平均分辨率约为 90kb
DNA 序列图	染色体 DNA 的完整核苷酸序列	1bp

尽管分辨率比较粗略，但细胞遗传图为通过原位杂交排列人类 DNA 序列提供了一个非常有用的基本框架，并且明确的细胞遗传断裂点可能成为另外的作图工具。利用罕见-切割限制性核酸酶，也产生了**长范围限制性酶切图**（long range restriction map），如绘制了 21q 上 *Not* Ⅰ 限制性酶切图（Ichikawa *et al*.,1993）。但是在 HGP 中，首张人类基因组高分辨率物理图可能是基于细胞 DNA 克隆技术构建基因组 DNA 克隆文库而绘制的（基本原理见节 5.3.4）。一旦可行，就可以筛查这些文库来鉴定单个的克隆，随后克隆被分组至含有来源于同一染色体和染色体亚区的插入子的克隆组中。最后，才有可能将克隆组成跨越大的染色体亚区乃至最终整个染色体的组。

构建基因组 DNA 克隆文库，通常始动于一个永生化细胞系（淋巴母细胞），以便提供不断更新的同质细胞群来源。采用标准方法从细胞系中分离基因组 DNA 之后，DNA 经限制性核酸酶切割并克隆到一个合适的载体中以便构建选择的基因组 DNA 文库。早期的尝试一般使用 λ 噬菌体和黏粒载体来构建插入片段在 15～40kb 范围的文库（节 5.4.2）。对克隆插入片段进行筛查（以前是通过与事先分离的小 cDNA 克隆进行杂交，但现在通过 PCR），然后通过**染色体原位杂交**（chromosome *in situ* hybridization）技术很容易将其定位于染色体亚区（图 2.16、图 2.17；注：通过与分裂中期染色体杂交来定位更短的 cDNA 克隆，在技术上比定位长的基因组克隆更困难，通常并不这样做）。

在早期的人类基因组 DNA 文库中，绝大部分克隆既没有被命名（因为它们插入片段 DNA 的身份是未知的）也没有被定位。逐渐地越来越多的 DNA 克隆被定性，使得相应的（同源的）基因组 DNA 克隆被鉴定并随后定位于染色体亚区（cDNA 克隆更容易定性，因为 cDNA 克隆插入片段短，测序相对较快，并且文库不很复杂，特定类型的克隆由于差异基因表达而通常占有优势，如血样中的珠蛋白转录物）。

为帮助人类基因组 DNA 克隆定位，还开发了一些其他的方法。它们包括：

- **富集启动 DNA**。使用与 FACS 细胞分类器分选细胞的相同原理，通过**流式细胞仪**（flow cytometry）来纯化各个染色体，来代替使用整个基因组 DNA。收集足够数目的特定类型的染色体，就可以构建**染色体特异性 DNA 文库**（chromosome-specific DNA library）（Davies *et al*.,1981）。额外的**染色体微切割**（chromosome microdissection）技术使人们可以利用从特定染色体亚区分离的 DNA 构建 DNA 文库（Ludecke *et al*.,1989）；
- **杂种细胞定位**（hybrid cell mapping）。对基因组克隆末端的一小段进行测序之后，

就可以对这段特异序列设计 PCR 分析，随后用于对缺少特定染色体或染色体亚区的杂交细胞嵌板进行分型（框 8.4）。

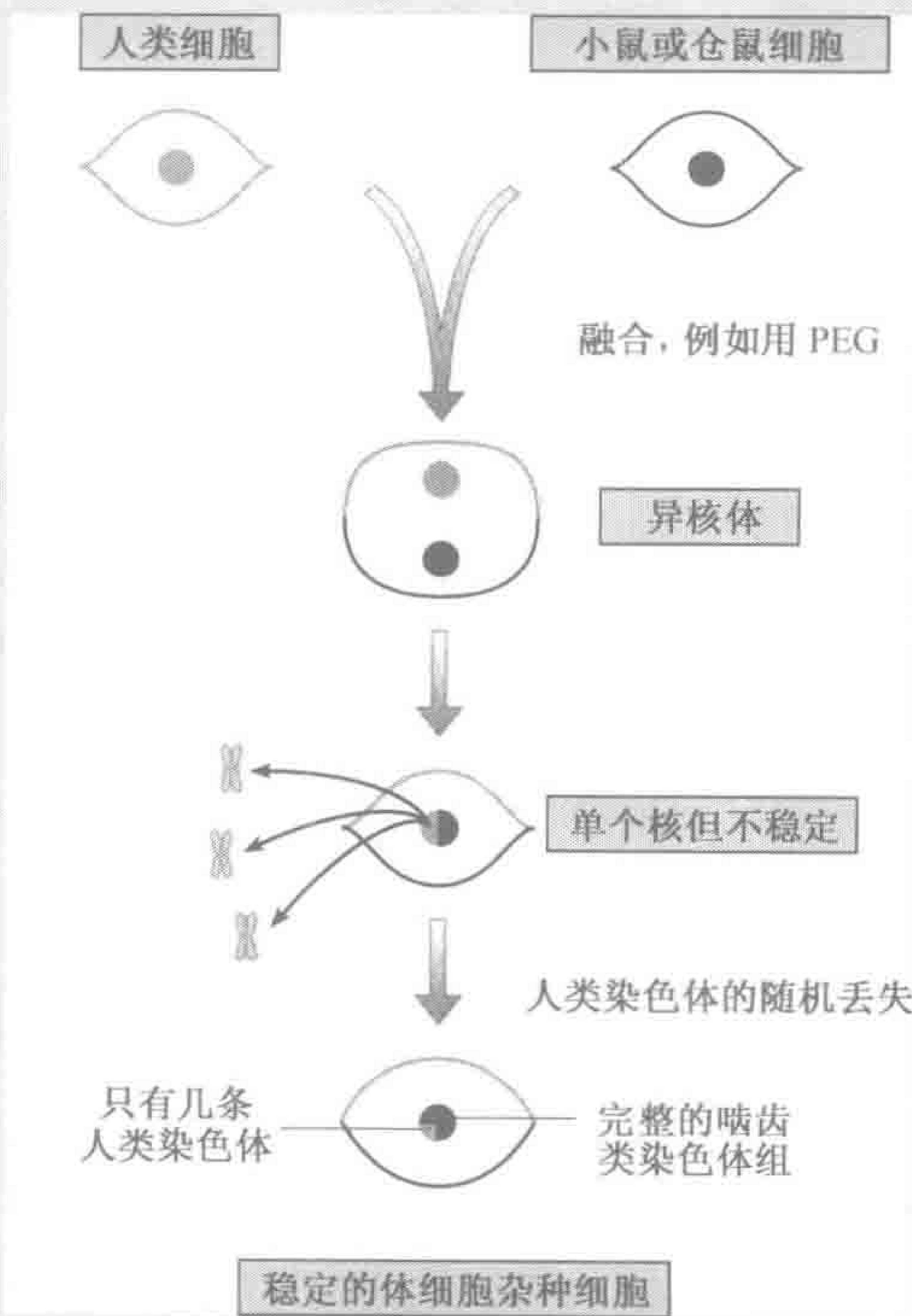
▶通过上述各种定位技术之一对已预先定位于物理图中的 DNA 片段进行**遗传连锁**（genetic linkage）。

框 8.4　杂种细胞作图

体细胞杂种的基本原理

在特定实验条件下，可以诱导不同物种的培养细胞相互融合，形成**体细胞杂种**（somatic cell hybrid）。为了人类定位的目的，通常把人类细胞与啮齿类动物（小鼠或仓鼠）细胞融合构建杂种细胞。最初的融合产物称为**异核体**（heterokaryon），因为细胞含有人类细胞核和啮齿类细胞核。最后异核体进行有丝分裂，两个核膜溶解。因此人类和啮齿类染色体汇合成一个细胞核。这样的杂种细胞是不稳定的。由于未知的原因，大部分人类染色体在接下来的细胞分裂周期中不能复制并丢失。这样最终就得到各种或多或少稳定的杂种细胞系，每个具有整套的啮齿类染色体和几种人类染色体（见图，上板）。人类染色体的丢失本质上是随机发生的，但可以通过选择加以控制。

携带不同人类染色体的杂种细胞嵌板可用于将一个人类基因或 DNA 序列定位到某条特异性人类染色体上。但是，使用**单染色体杂种**（monochromosomal hybrid）（只含有一种人类染色体的细胞）嵌板是最有效的，总共存在全部 24 种人类染色体（Cuthbert *et al.*,1995）。为了制作单染色体杂种细胞，人类供体细胞必须用秋水仙胺处理，使整套染色体变成分散的核下小体（**微核**，micronuclei）。然后对其进行离心可以形成**微细胞**（microcell）——包含有一个微核、一薄层胞浆并由一完整的胞膜包被。该微细胞与啮齿类受体细胞融合（**微细胞融合**，microcell fusion），形成杂种细胞，其中有些具有单个人类染色体（例如，Warburton *et al.*,1990）。



体细胞杂种原理

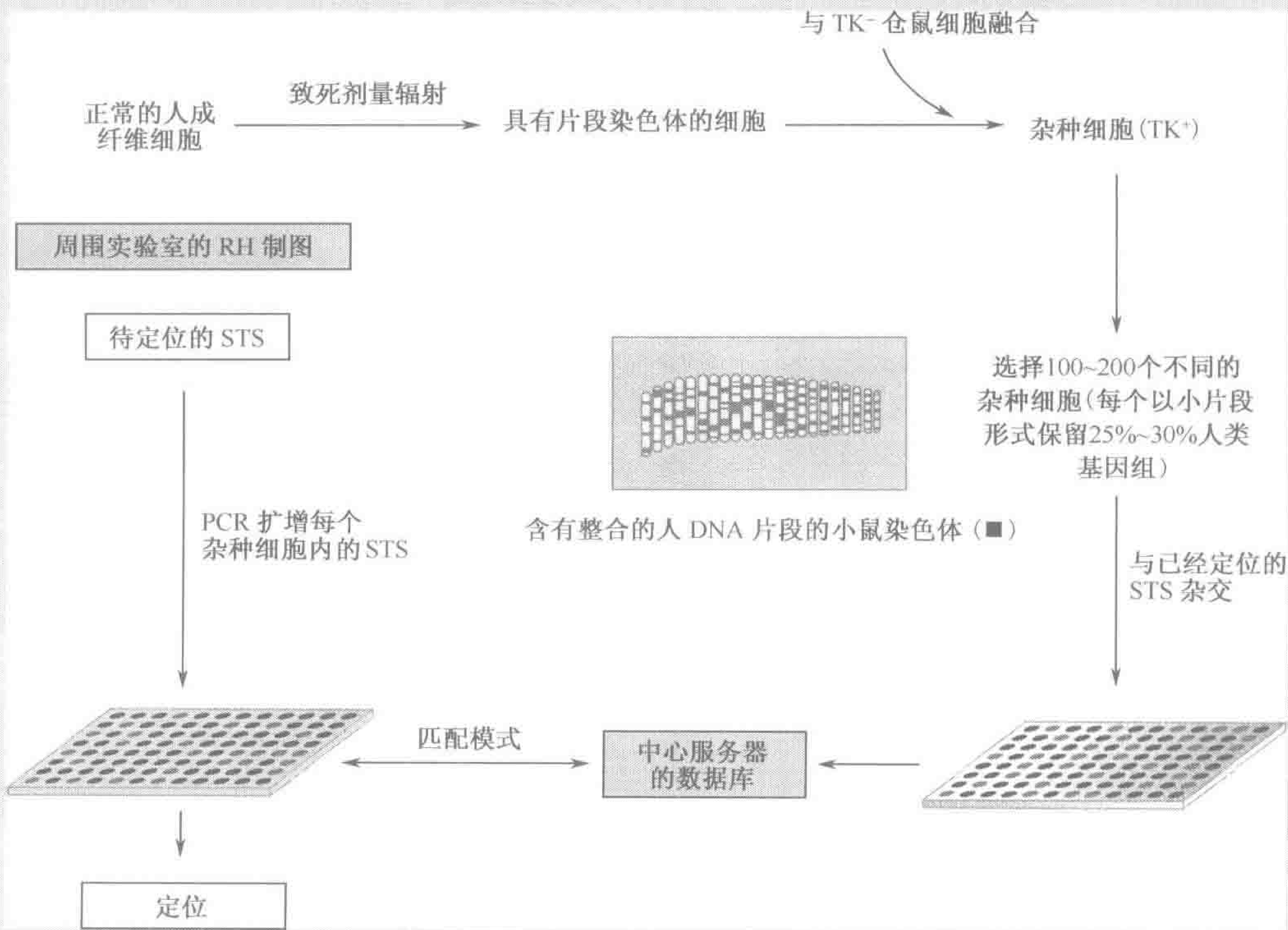
框 8.4　杂种细胞作图（续）

辐射杂种（radiation hybrid）

利用**辐射杂种**嵌板，人类基因或DNA序列也可以定位到染色体亚带，杂种细胞包含有人类染色体片段和全套啮齿类染色体。给予供体细胞致死辐射剂量，引起染色体断裂就可以得到人类染色体片段（平均片段大小是辐射剂量的函数）。辐射之后，供体细胞与啮齿类受体细胞融合，使用一种选择系统来挑选已接受一些供体染色体片段的受体细胞（Walter *et al.*,1994）。

以前使用的**单染色体辐射杂种**（monochromosomal radiation hybrid）嵌板，其供体细胞系是单染色体杂种细胞：在这种情况下一种人类染色体片段和碎裂的啮齿类染色体片段都整合到啮齿类供体细胞的染色体中（Cox *et al.*,1990）。现在它们被**全基因组辐射杂种**（whole-genome radiation hybrid）嵌板所代替，其供体是辐射处理的正常人类二倍体细胞（Walter *et al.*,1994）。对任一个杂种细胞来说，只有一部分碎裂的人类染色体片段会发生整合，这样不同的杂种细胞就有不同的人类基因组部分整合到啮齿类受体染色体组中。

用一套人类DNA标记把杂种嵌板分型就可以制成辐射杂种（RH）细胞图。由于各个杂种细胞具有不同但又重叠的人类基因组，因此不同的标记可能出现在一些杂种细胞中，而在另一些杂种细胞中不存在。虽然大多数片段整合的方式是随机的，但如果各个标记定位于特定染色体上的邻近位置，那么它们就可能产生相关的分型方式。辐射杂种图的原理是减数分裂连锁分析的再现（13章）：同一染色体上两个DNA序列越近，它们被二者之间断裂点分开的可能性就越小。两个标记之间染色体断裂的频率用 θ 值确定，类似于减数分裂定位的重组频率。θ 值从0（两个标记从不分开）到1.0（两个标记总是分开）变化。



辐射杂种定位

框 8.4　杂种细胞作图（续）

就像在减数分裂定位中一样，θ 值低估了在同一染色体上相隔较远的两个标记之间的距离，这是因为一个细胞可以接受分开片段的两个标记。RH 作图函数提供一个更精确的估计值，$D=-\ln(1-\theta)$，与减数分裂连锁分析使用的 Haldane 作图函数类似（节 13. 1）。D 用厘伦琴（cR）表示。D 取决于辐射剂量，所以与拉德（rads）数负相关。比如，两个标记间距为 1 cR_{8000}，表示它们暴露于 8000 拉德的 X 线后有 1%的断裂频率。

人类基因组计划中两种辐射杂种细胞嵌板特别重要。基因桥 4 嵌板含有 93 个人类－仓鼠辐射杂种细胞，人类片段平均大小为 25Mb，每个杂种细胞保留 32%的特定人类序列。实验室通过记录 93 个 Genebridge 杂种细胞的模式并将其与中心服务器拥有的以前定位的标记模式进行比较可以定位任一未知 STS（见图，下板）。另一种人类－仓鼠嵌板——Stanford G3 嵌板——是用高剂量辐射制成的，所以人类片段的平均长度要小一些。G3 中的 83 个杂种细胞平均保留 16%的人类基因组，平均片段长度 2. 4 Mb。因此 G3 可用于更精确的定位。大规模应用这些嵌板得到的重要成果见 http://www.ncbi.nlm.nih.gov/genemap98/（Deloukas *et al.*,1998）。

基于 YAC 的人类基因组物理图

在 1990 年人类基因组计划正式开始时，可用的基因组 DNA 文库含有最长不超过 40kb 的插入片段（黏粒克隆），其中绝大部分是没有命名的，也没有定位。由于人类基因组非常庞大，所以平均约 40kb 插入片段大小的黏粒文库可能需要几十万个不同的克隆来确保接近 100%基因组出现于文库中的高可能性。筛查这些复杂的文库把各个克隆分开，并把这些克隆分成相关的各组，是一项令人生畏的任务。

为减少筛查大量克隆带来的问题并且易于将它们分成相关的组，提供非常大的插入片段大小的克隆系统是很吸引人的。为了达到这个目的，人们通过制备人工真核染色体开发了新方法。这个染色体系统建立在线性酵母染色体基础上，已知酵母染色体只有非常小的序列但对其功能是不可缺少的。通过纯化这些序列并把它们与人类 DNA 的大片段连接在一起，就有可能制作含有 Mb 大小人类插入片段，且在酵母细胞中仍表现为染色体的杂种分子，即所谓的**酵母人工染色体**（yeast artificial chromosome，YAC）一详见节 5. 4. 4。

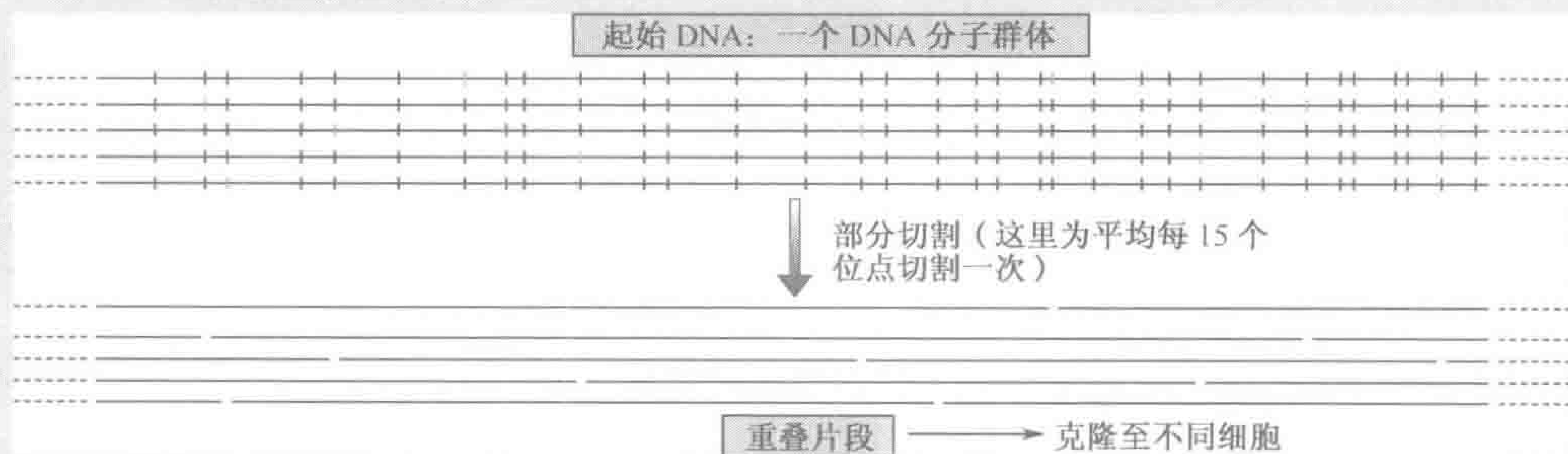
YAC 文库，其平均插入片段大小为 1Mb，只需 12000～15000 左右个克隆就可以很好地表示人基因组，并具有能够使大的基因（及基因簇或其他功能片段）包含于单个克隆的优势。用这种方法，巴黎 CEPH 实验室的 Daniel Cohen 及其同事们构建了人类基因组的 YAC 物理图，它是第一张非常详细的高分辨率物理图（Cohen，*et al.*,1993）。后来这个团队又公布了一个更新了的 YAC 图谱，涵盖了大约人类基因组的 75%，由 225 个平均大小为 10Mb 的叠连群组成（Chumakov *et al.*,1995）。

YAC 图（以及所有其他基于克隆的物理图）的根本原理是根据插入 DNA 片段的染色体亚区来源将文库中各克隆进行排序。这个图是通过确定克隆组而建立的，在克隆组中，插入的 DNA 片段来源于共同的染色体亚区，并且任何一个克隆中的插入 DNA 与克隆组中一些其他克隆的插入 DNA 重叠。这意味着相关的染色体亚区是由部分重叠的克隆线性排列所表示的，而没有留下任何缝隙。这样一组连续的克隆 DNA 序列称为**克隆叠连群**（clone contig）（框 8. 5）。

框 8.5　通过建立克隆叠连群进行物理作图

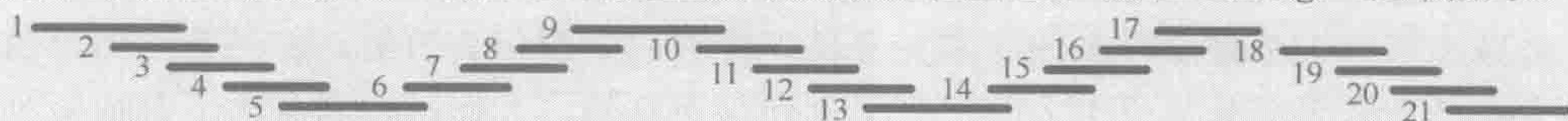
基因组 DNA 文库的构建涉及仔细控制克隆 DNA 的部分限制性消化。常用于从起始 DNA 产生限制性片段的酶是 *MboI*，它可以识别 4bp 的识别序列（↓GATC），平均每 300bp 左右就可以切割 DNA 一次。把起始 DNA 暴露于极低浓度的酶中并作用很短时间，可用的全部限制酶位点只有很少的位点被切割。例如，当制作 BAC 文库时，所需的克隆片段大小大约 200kb，为了实现这一目标只需要切割不到 1%的 *MboI* 位点。

起始 DNA 通常从几百万个二倍体细胞获得，因此每个原始的 23 种人类 DNA 分子（对应 23 种不同的染色体）应该是由几百万个原始染色体 DNA 分子的相同拷贝组成的。但是，当进行部分限制性消化时，DNA 分子被随机切割。因此对于任何一个染色体亚区，几百万左右的相关分子都表现出不同的限制性位点切割模式（见图）。



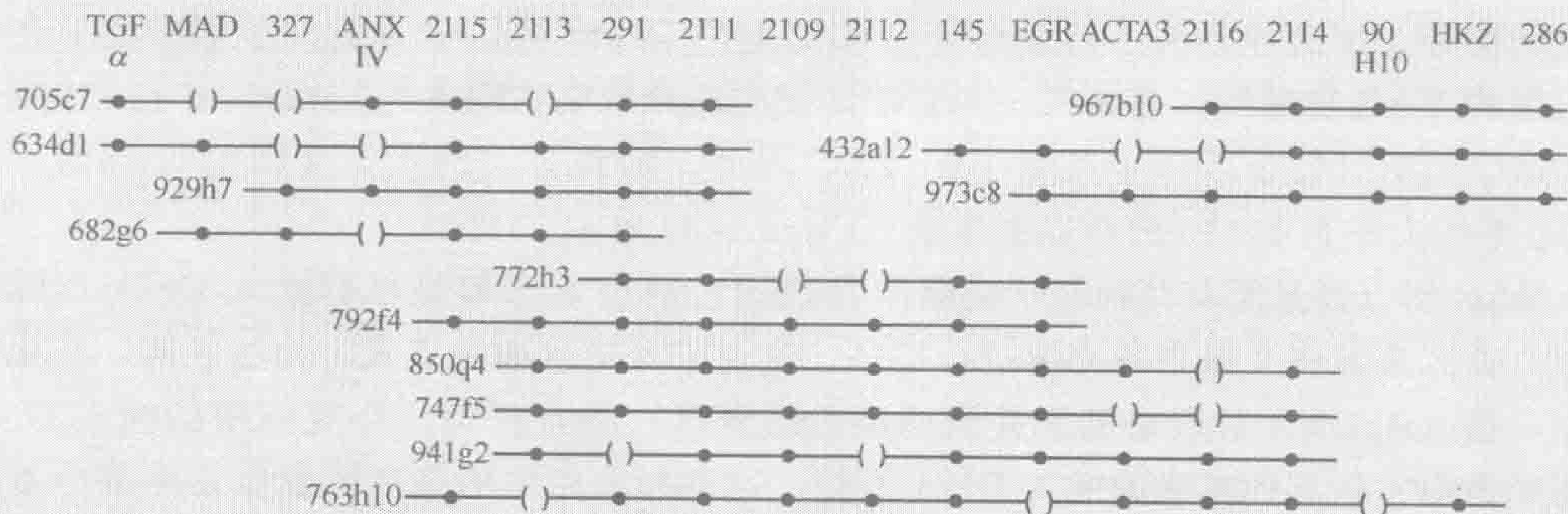
上图：通过部分限制性切割产生重叠 DNA 片段

部分限制性消化产生了不同的片段，这些片段拥有来自同一染色体亚区的重叠序列。尽管各个片段将被克隆到不同的宿主细胞中，仍有可能通过筛查不同克隆插入片段之间的相似之处，鉴定具有重叠插入片段的克隆，最后确定一组有重叠插入片段的克隆；结果在这组克隆的插入片段中出现染色体区域内所有原始 DNA 序列而没有缝隙，即所谓的**克隆叠连群**（clone conting）（见中图）。



中图：克隆叠连群-具有部分重叠 DNA 插入片段的系列克隆

具有重叠插入片段的克隆可以用不同的方法来鉴定。对于人类 YAC 图来说（节 8.3.2），重叠克隆鉴定是利用**克隆-克隆杂交法**（clone-clone hybridization）（用源自一个 YAC 的探针去杂交其他 YAC 克隆）或**克隆指纹法**（clone fingerprinting）（比较克隆来观察它们是否共享特征性的特定重复序列或限制性位点的空间模式）。最近，**STS 容量绘图**（STS-content mapping）成为更受欢迎的方法，就像 Hudson 等（1995）公布的 STS 图中所描述的。这里所有的克隆根据一组**序列标签位点**（sequence tagged site）标记的每一个标记的存在进行分型，然后再根据克隆是否共有两个或更多特定的 STS 标记进行分组（见下图）。



下图：2 号染色体上一个 STS 叠连群的例子

人类基因组高分辨率的序列标签位点图

克隆叠连群图的准确性关键依赖于克隆插入 DNA 能真正代表原始基因组序列的程度。虽然人类 YAC 图谱是一项伟大的成就，但相当大的基因组区域没有在图谱中表示出来，而且存在一个主要的内在缺陷：插入 DNA 通常不能真实地代表基因组 DNA。大的 YAC 插入片段倾向于生重排（包括内部序列的丢失），还有一个实质性问题就是**嵌合状态**（chimerism）（单个转化细胞含有两个或更多来自基因组不连续部位的通常是来自不同的染色体，人类 DNA 片段，是共连接或共转化的结果－分别见图 5.5A 和图 5.5B）。

为了避免因克隆插入片段失真而产生的问题，HGP 物理作图策略也强调了发展基于**序列标签位点**（sequence tagged site，STS）作图的必要性（框 5.4）。拥有足够高密度的 STS 界标，YAC 文库中插入片段不稳定的问题就可以避开：大量的 STS 意味着，通过其他类型克隆（BAC，P1，PAC 等）的 STS 分型可以快速地恢复任何问题区域的物理学覆盖。利用这种方法，Whitehead 生物医学研究所和马萨诸塞技术研究所的 Eric Lander 及其同事们报道了人类 STS 图界标的完成，它含有超过 15000 个 STS，平均间隔少于 200kb（Hudson *et al.*,1995）。

人类 STS 图是一种整合物理图，在此图中，STS 用于区分：(a) 人类**辐射杂种细胞**（radiation hybrid cell）嵌板（框 8.4）；(b) CEPH YAC 文库。STS 标记有两种类型，非多态性的和多态性的。非多态性的 STS 标记包括，随机测序基因组克隆然后在非重复区域设计 PCR 引物而得到的 STS，以及从 cDNA 序列中选择得到的 STS（使用与未被内含子分割的序列一致的引物），即所谓的**表达序列标签**（expressed sequence tag，EST）（节 8.3.4）。多态性 STS 标记大多数由微卫星标记组成，其中大部分已被 Généthon 团队用于人类遗传作图。

Hudson 等（1995）公布的 STS 图为人类基因组提供了一个广泛的物理框架，因其含有超过 2400 个 EST，所以也提供了一个胚胎的人类基因图。从现在开始，人类基因组计划的焦点转到两个方向上：制作高分辨率的基因（转录物）图；利用已有的 STS 图为应用细菌人工染色体克隆构建克隆叠连群提供框架。

8.3.3　人类基因组计划的最后阶段关键取决于 BAC/PAC 克隆叠连群

测序策略

由于 YAC 插入片段通常不能忠实地反映原始的起始 DNA，所以需要第二代人类基因组克隆叠连群图提供可以进行测序的克隆 DNA。人们选用 BAC 和 **P1 人工染色体**（P1 artificial chromosomes，PAC）作为克隆系统，因为尽管其插入片段大小（100～250kb）比 YAC 插入片段大小要小得多，但更大的插入片段保真性远远超过了这个缺点（节 5.4.3）。因此各种不同的人类 BAC 或者范围较小的 PAC 文库，都被用作测序的物理模板。

由政府出资的人类基因组计划所采用的测序策略是建立在等级式鸟枪法测序基础之上（而 Celera 使用了全基因组鸟枪法测序：见图 8.3）。鸟枪法测序是指一般通过超声

把起始 DNA 随机分割成小片段，接着进行末端修复，然后把这些片段克隆至载体，就可以很容易制成单链重组 DNA 并直接用来测序（框 7.1）。在等级式鸟枪法测序方法中进行鸟枪法测序的 DNA 是已准确定位于物理图的各个 BAC 克隆已纯化的插入片段；而全基因组鸟枪法测序设计涉及直接对分离的基因组的 DNA 进行鸟枪法测序（图 8.3）。

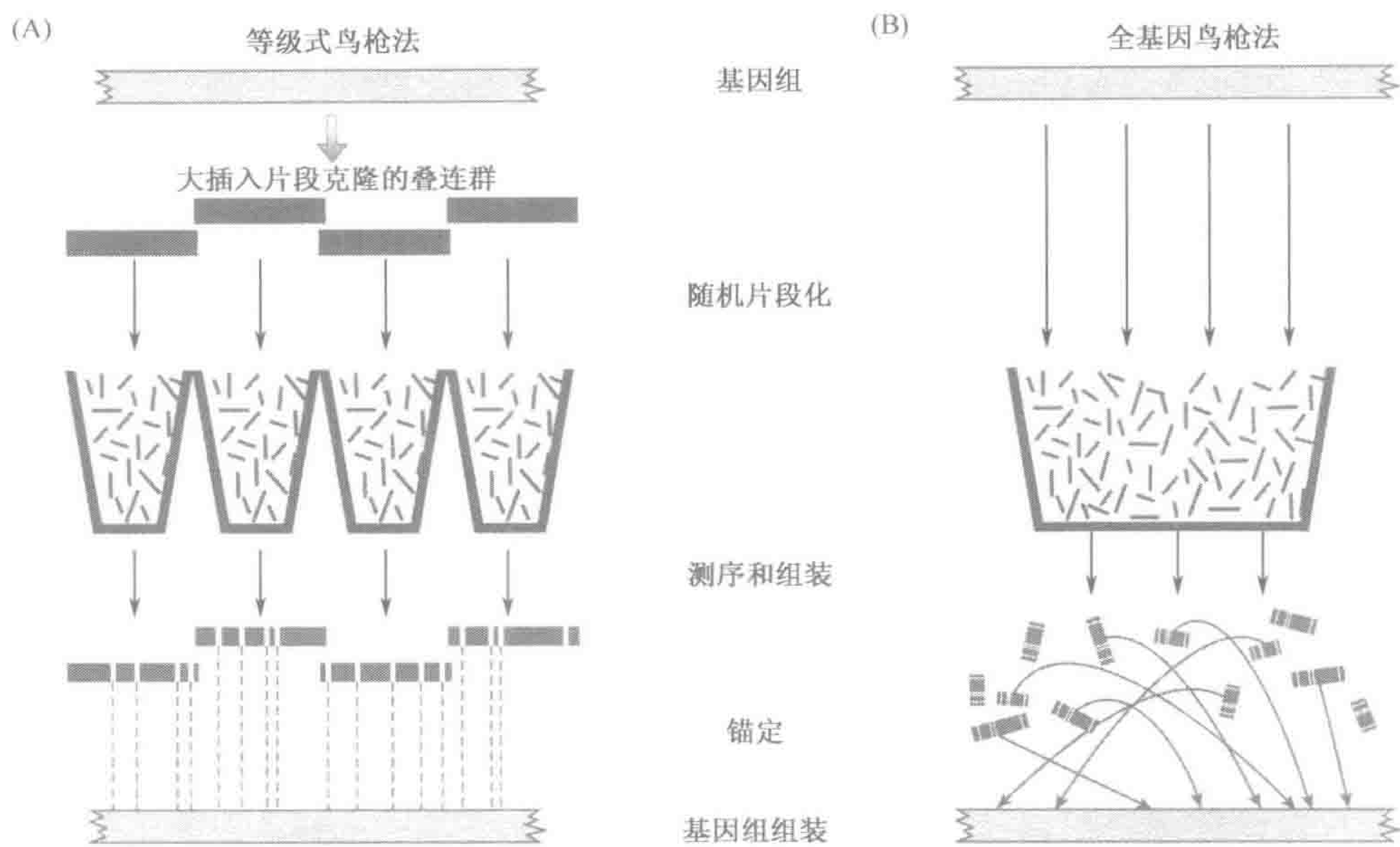


图 8.3　人类基因组测序的不同鸟枪法测序策略

（A）**等级式鸟枪法测序**（hierarchical shotgun sequencing）。通过部分限制性消化使人类基因组 DNA 片段化，得到的大限制片段克隆至 BAC 载体制作 BAC 文库。利用 STS 标记对所有克隆进行分类，鉴别出具有重叠插入片段的克隆，从而把 BAC 克隆组成大的叠连群。选定的 BAC 克隆插入片段进行鸟枪法克隆和测序。然后把 BAC 的测序片段集合起来产生 BAC 序列，再移去重叠部分就得到整个 BAC 序列。（B）**全基因组鸟枪法测序**（whole genome shotgun sequencing）。在这里分离的基因组 DNA 直接进行鸟枪法克隆和测序，然后测序的片段被集合成跨越 Mb 的大叠连群。经 the National Academy of Sciences，USA 允许，改编自 Waterston 等.（2002）. Proc. Natl Acad. Sci. USA 99，p. 2713。

人类基因测序所使用的测序方法学的基础是 25 年前 Fred Sanger 及其同事们发明的双脱氧测序法。虽然基本方法没有改变，但是序列生成和数据分析方面的自动化使得效率有了很大提高。荧光标记自动 DNA 测序仪和后来毛细管测序仪的发明（节 7.1.2）使得测序能力大大提高。各种专用的计算机程序有助于序列翻译和组装，特别是 PHRED（分析原始的序列迹线并在每个碱基位置提供一个质量评分来显示指定的碱基信号是正确的可信度）和 PHRAP（通过筛查两个或更多独立的鸟枪法克隆之间共享的重叠序列，把原始序列组装至序列叠连群）。

重复 DNA 的问题

组装各个克隆序列并找出重叠部分（由此重新构建基因组序列）关键取决于一个重

要的假设：重叠的序列在基因组中是唯一有的。但是人类基因组的很大部分（大约 50%）是由重复 DNA 组成的（节 9.4 和 9.5）。因此在克隆序列中寻找有意义重叠的证据时，要尽可能避免那些已知的高度重复散在 DNA 类型，如著名的 LINE-1 和 Alu 重复。

除了上述的预防措施，那些富含已知重复序列的区域也存在问题，而另一个担忧是先前未鉴定的低拷贝数重复（这是一个非常真实的问题，因为人们随后发现大部分基因组经历节段性复制，序列跨越在基因组的两个或更多区域中出现的几十个 kb 有时是几百个 kb——节 12.2.5）。至少在由政府出资的 HGP 中等级式鸟枪法克隆策略促进了序列组装；而私人 Celera 公司使用的全基因鸟枪法克隆策略对计算能力有很高的要求，批评家争论认为，这种策略如果单独使用的话，则必定会由于人类基因组的复杂性和高度重复含量而失败（框 8.6）。

框 8.6　基因组计划中的合作、竞争和争议

由政府出资的基因组计划中的合作和竞争

由于它们的规模，所以基因组计划是大规模的任务。在许多情况下有许多值得称赞的合作范例：不同的中心和实验室共享资源，一致细分任务等。酵母基因组计划就是一个很好的例子，涉及不同的欧洲中心之间高度有序的合作，并一直持续到功能分析。而有些情况，不同实验室之间的激烈竞争造成的紧张十分明显，导致了无用的重复工作，例如大肠杆菌基因组计划。在这项计划中美国和日本的团队进行了竞争，最后的结果几乎相同：美国实验室比日本实验室早一周把序列存入基因库中。

公共和私人团体之间的紧张关系：基因专利

公共和私人团体的不同目的给基因组计划造成了多种压力。早期的争论领域就是关于**基因专利**（gene patent）。这个问题最早出现在 1991 年，美国 NIH 申请了 7000 多个人脑 cDNA 克隆片段的专利，而这些序列已作为 Craig Venter 博士领导的一项 EST 作图试验的一部分已建立了。这一企图遭到了科学界的广泛反对，尤其是这些表达序列的功能还一无所知。在压力下，美国专利局拒绝了这个申请。另一个重要的新问题第一次被提出－谁拥有人类基因组？（Thomas *et al.*，1996），对我们非常简单的遗传财产进行商业垄断的观点对许多人来说都是烦扰和厌恶的。

接下来 Venter 博士离开了 NIH，成立了一个新的商业化支持的研究所－基因组研究所，采用一种工业化模式的方法进行 EST 测序，并迅速编制成世界最大的人类基因数据库。1994 年 4 月，SmithKline Beecham 药品公司投资八千万英镑作为 Venter 数据库的唯一股份，并宣布任何想使用它们的科学家必须将他们专利性发现转让给该公司。一个公司试图垄断大部分表达人类基因组的企图再次震惊了科学界。许多人认为，只有在鉴别出基因的功能之后才可以申请专利，而不是鉴别之前。现在已经对已知与某些功能相关的人类 DNA 序列授予了几千个专利（Thomas *et al.*，1996），但 1998 年 10 月美国专利局授予了第一个 EST 序列专利（对 Incyte 药品公司）。如 Knoppers 所叙（1999），人类基因组专利性的问题持续存在着。

公共和私人团体之间的紧张关系：基因组测序

私人赞助的基因组测序也存在很多争议。Celera 人类和小鼠基因组测序的工作是在与建立时间更长的由政府出资的计划的激烈竞争中完成的。1999 年，Celera 大胆地宣布他们的**全基因组鸟枪法策略**（正文和图 8.3）可以在两年内制作出人类基因组序列的草图，很快就能赶上由政府出资的 HGP 缓慢的逐个克隆进行作图和测序的策略。

框 8.6　基因组计划中的合作、竞争和争议（续）

碰巧，HGP 和 Celera 同时公布了人类基因组序列草图的完成（国际人类基因组测序协作组，2001 年；Venter *et al.*,2001）。但是，这绝不是一个公平的竞争，因为自始至终 Celera 拒绝容易和免费获得他们的数据（拒绝外界了解他们的数据，即使交纳了所需的昂贵费用，仍然受到限制）。与此完全相反，由政府出资的实验室致力于即时广泛地传播他们新的序列数据（每 24 小时在互联网上更新一次）。和其他人一样，Celera 可以持续地没有约束地获取由政府出资的 HGP 的 DNA 序列数据，并且无耻的截取其中大块段的公共获得的人类基因组序列数据，再处理后据为已有。结果 Celera 测序工作极大依赖于公共项目产生的数据。Celera 使用的是非常有难度的全基因组鸟枪法策略，因此很多人都怀疑 Venter 等（2001）报道的人类基因组序列是否可以完全证明全基因组鸟枪法策略可行。反而，Waterston 等（2002）提供数据证明 Venter 等（2001）报道的 Celera 序列根本不是一个独立的人类基因组序列（因为许多 Celera 序列数据都是通过截取和重组大量 HGP 序列得到的）。

8.3.4　第一个高密度人类基因图谱是根据 EST 标记建立的

在 HGP 刚开始时人们曾争论过，是进行彻底的研究（无差别的对全部三十亿碱基进行测序），还是最初仅集中于代表编码 DNA 序列的非常小的部分，这部分至少是最令人感兴趣的和医疗相关的部分。最终全基因组测序的支持者在辩论中胜出，他们强调找到全部基因是很困难的（有些基因的表达非常受限），而且有些非编码 DNA 在功能上很重要，例如对于染色体功能很重要的调节元件和序列。然而在计划初期，竞争性商业利益却优先放在寻找基因上。

最初的方法涉及 cDNA 克隆 3′非翻译区短序列的大规模测序，这些克隆是从各种人类 cDNA 文库中随机选择的序列（Adams *et al.*,1991）。这些短序列称为**表达序列标签**（expressed sequence tag，EST），因为就像更常用的**序列标签位点**（sequence tagged site）一样，可以对这些序列设计一特定的 PCR 实验来找出表达序列（特定的 3′UTR 序列的基因组序列比编码 DNA 通常较少被内含子分隔，所以 3′UTR EST 的 PCR 引物常可扩增基因组 DNA 样本中特定的序列）。后来测序延伸到 cDNA 克隆的 5′端，最终从几十万个人类 cDNA 克隆的末端获得序列。

获得大量的人类 EST 之后（私人和由政府出资的研究小组共同完成），下一个任务就是把它们安置到人类基因组的物理图上。这需要根据各个 EST 把 YAC 叠连群分类，或筛查人类－仓鼠**辐射杂种**（radiation hybrid）细胞嵌板（辐射杂种细胞原理见节 8.3.2 和框 8.4）。之前通过杂交研究已经得到人类基因粗略的染色体分布（图 8.4），但把大量的 EST 放置到物理图上是构建人类基因图的第一个系统方法。

整合与人类遗传图相关的作图数据，并与染色体的细胞遗传带型图互相参考。为了确定必要的基因组，人们尝试整合不同 EST 的作图信息，就像 UniGene 系统一样（http：//www.ncbi.nlm.nih.gov/UniGene/）。首批相当广泛的人类基因图是 Schuler（1996）和 Deloukas（1998）报道的。后者报道的图谱位点估计有 30000 个（不准确的），当时认为这还不到全部人类基因目录的一半。但是人类基因组的测序工作却带给人们一个意想不到的惊奇（见下节）。

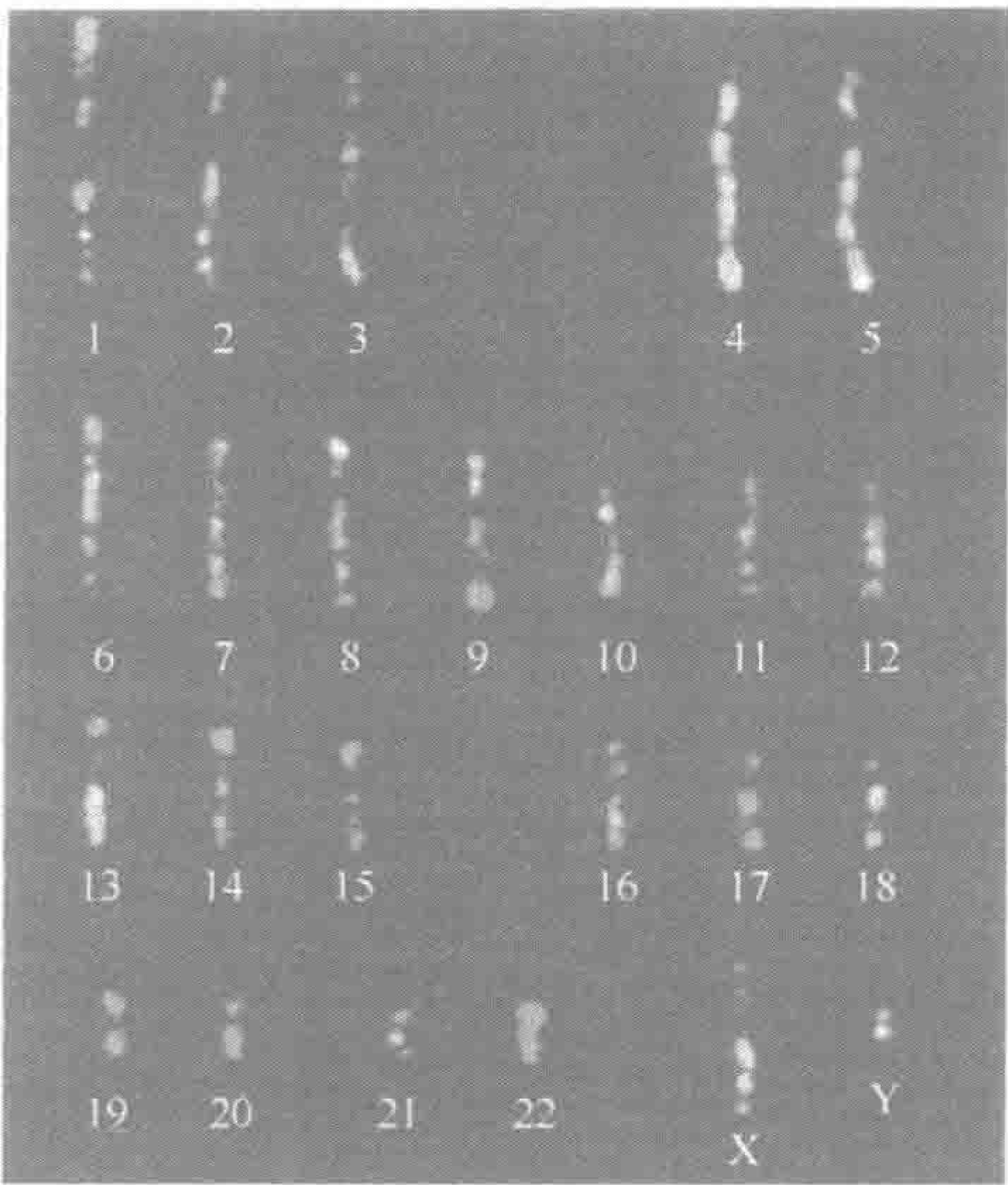


图 8.4　一张早期的全基因组人类基因分布图

大多数人类基因都与 CpG 岛相关（框 9.3）。纯化的人类 CpG 岛片段用德克萨斯（Texas）红染料标记，并与人类中期染色体进行杂交（Craig 和 Bickmore，1994）。晚期复制的染色体区域（大多数转录失活）显示为绿色［掺入异硫氰酸荧光素（FITC）标记的溴脱氧尿苷（BrdU）的结果］。黄色区域（红色和绿色信号重叠）表示寡合基因的（或严格来说是 CpG 岛）晚期复制区域。基因组中早期复制的基因缺乏区域是看不到的（因为没有复染），就像着丝点一样（那里没有抗 BrdU）。某些染色体（如 22 号染色体）基因密度高（由标记 CpG 岛片段的红色表示），而其他染色体（如 4、18、X、Y 号染色体）基因较少。经 Nature Publishing Group 允许，摘自 Craig 和 Bickmore (1994). Nature Genet. 7，376－381，图 1。

8.3.5　人类基因组序列草图提示有 30000～35000 个基因，但获得一个精确的总数却很困难

在人类基因组测序之前，人类基因总数的预测大多在 60000～100000 个（是建立在从各种有限的数据集中推断的基础上）。在 2001 年报道序列草图后，修正的估计数目惊人地低，也许只有 30000～35000 个基因。仅仅比秀丽新小杆线虫——一种 1mm 长，只有 959 个体细胞的蠕虫——多 50％的基因，人们又对人类价值产生了疑问，也许比长期建立的 C **值谬论**（C-value paradox）更令人吃惊（细胞 DNA 含量并不总是与功能的复杂性相关；有些类型阿米巴虫每个细胞含有的 DNA 比我们还要多得多——见表 3.1）。

确定人类基因准确数目的困难

即使到 2003 年，全部的人类基因组都基本测序完毕，人类基因的具体数目仍无法确定，尽管最近的估计（Ensembl build 29）提示人类基因总数接近 30000 个或者也许更少（见小鼠基因组测序协作组 2002 年发表的文章中人－小鼠比较）。考虑基因是如何定义的对于了解估计精确基因数目所存在的困难是有启发性的。如节 7.2 详述的那样只

有两个必要的基因特异性标准：转录成为 RNA 和进化过程中保守序列的证据。然而，在实际水平上，通过实验方法，根据任一标准来鉴定全部的基因都是非常困难的，因为：

- 虽然寻找基因相关的转录序列通常是很有价值的（Camargo *et al*.,2002），但表达水平低和（或）只在特定细胞部位和发育阶段才表达的基因可能都没有很好的出现在可用的 cDNA 文库中。
- 缺乏大的可读框，很难鉴别那些编码非翻译 RNA 的基因。

由于上述原因，仅仅用实验方法通常丢失基因，当人类基因组序列草图公布时，实验支持存在的基因可能只有 11000 个，其他都是通过计算机推算出来的（电子克隆分析）。用于鉴定基因的计算机程序（Zhang，2002）把待测序列和其他已知基因序列比较（同源性筛查）并尝试着鉴定外显子（外显子预测程序）：

- **对序列数据库的同源性搜索**。对任何一个待测序列来说，核酸序列可与可用的数据库中所有核酸序列相比较，而且其可能的翻译多肽也可与所有已知的蛋白质序列相比较（表 8.2 和框 7.3）。随着越来越多的序列信息载入数据库，该方法成为鉴定在进化过程中高度保守的预示保守功能的序列的一个非常有效的途径；

表 8.2 作为核苷酸和蛋白质序列储存库的主要电子数据库

数据库类型	数据库	位置	URL
核苷酸序列	GenBank	由美国国立卫生研究院（NIH）的国家生物技术信息中心（NCBI）维护	http：//www. ncbi. nlm. nih. gov
	EMBL	由英国剑桥附近 Hinxton 欧洲生物信息学所（EBI）维护	http：//www. ebi. ac. uk
	DDBJ	由日本静冈县三岛市的国家遗传学研究所维护	http：//www. ddbj. nig. ac. jp/
蛋白质序列	SWISS-PROT	附有高质量注解的蛋白质序列。由瑞士日内瓦生物信息学研究所和英国休斯顿欧洲生物信息学研究所（EBI）合作维护	http：//ca. expasy. org/sprot http：//www. ebi. ac. uk/swissprot/
	TREMBL	EMBL 数据库中没有列入 Swiss-Prot 的编码序列的翻译产物	http：//www. ebi. ac. uk/tremble/ http：//ca. expasy. org/sprot
	PIR	由美国乔治镇国家生物医学研究基金会（NBRF）、日本国际蛋白质信息数据库（JIPID）和慕尼黑蛋白质序列信息中心（MIPS）合作维护。在世界许多地方都可使用	http：//pir. georgetown. edu/ http：//www. ddbj. nig. ac. jp/ http：//nips. gsf. de/ http：//www. hgmp. mrc. ac. uk/

- **外显子预测程序**。有些程序仅能利用输入序列的信息，如常用的 GENSCAN 程序（Burge and Karlin，1997）。其他一些程序则很大程度依赖于确定的同源性筛查。然而，到目前为止即使最好的程序如 GENESCAN，当用来检测已预先确定外显子结构的基因时，其鉴定外显子的成功率也仅为中等水平。物种交叉比较提供了另一个重要的鉴别外显子的方法，因为与大多数基因序列不同，外显子倾向于在进化中高度保守（Batzoglou *et al*.,2000）；

▶ **整合的基因寻找软件包**。人们制作了同时使用通用的序列同源性数据库搜索程序以及设计用于鉴定相关基序及外显子的程序的各种软件包。通常输出结果是图形格式的。常用的软件包有 *NIX*（http：//www. hgmp. mrc. ac. uk/Registered/Webapp/nix/；使用此软件包的输出结果见图 8.5）和 *Genotator* 程序（见http：//www. fruitfly. org/～nomi/genotator/genotator-paper. html）。

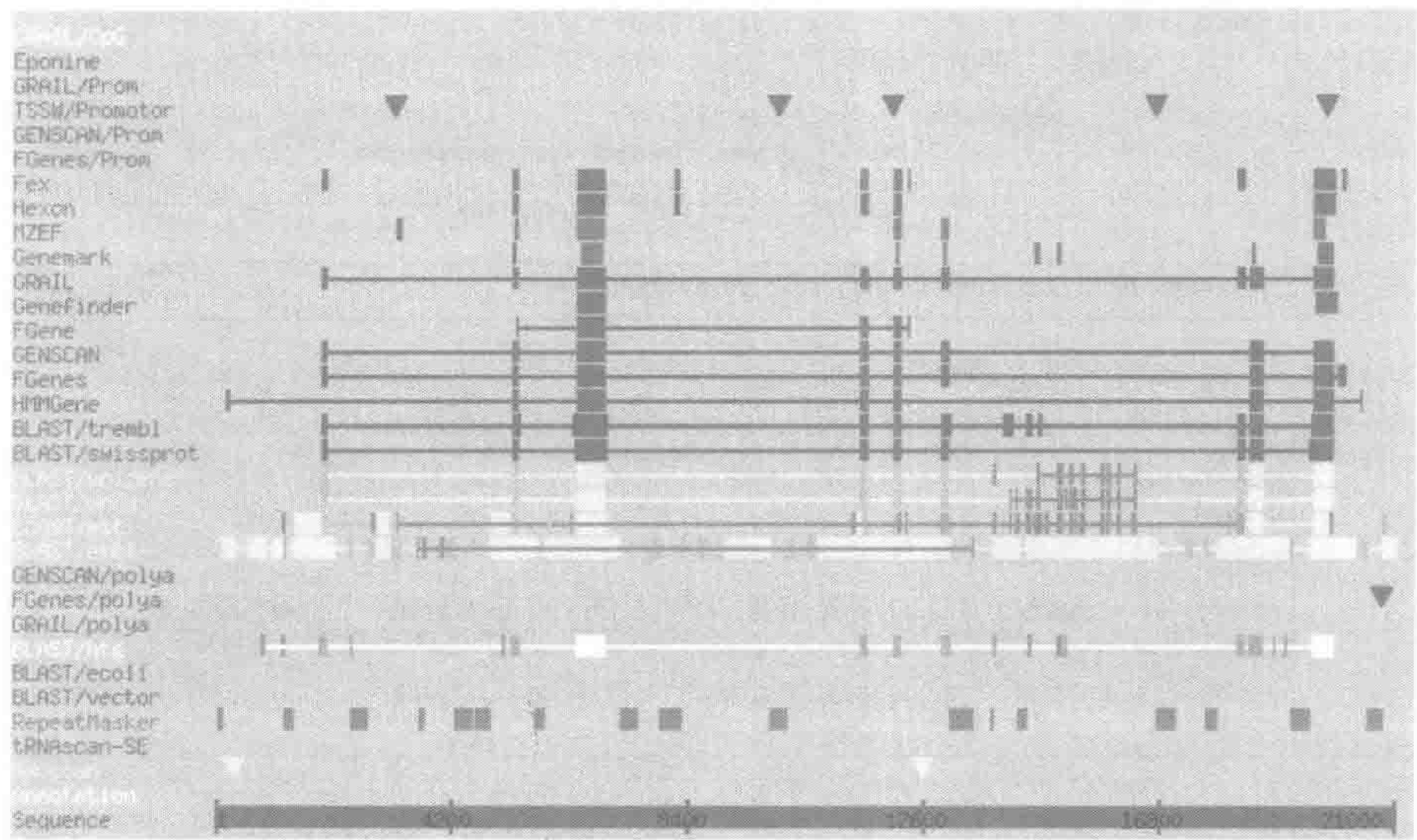


图 8.5　通过计算机分析基因组序列来查找基因

这个例子显示了对来自 12q24.1 上包括 Darier 病基因座的染色体区域的一段 PAC 序列的 NIX 分析（见文中）。该区域核苷酸的大小在底部用长条表示。分析包括应用寻找基因相关基序——如启动子序列（顶部绿色的反三角）、多腺苷酸化位点（赭色反三角）的程序，以及各种外显子预测程序（GRAIL、GENSCAN 等）。在各种 BLAST 程序中，与其他序列在核苷酸和蛋白质水平上明显的同源性序列用框表示。数据由英联邦 Newcastle upon Tyne 大学的 Victor Ruiz-Perez 和 Simon Carter 博士提供。

尽管用计算机推测基因和外显子取得一些成就，但在估计人类基因数量上仍存在问题，既有过度推测，也有推测不足。例如当一簇看似独立的基因结果证明是单个大基因的不同部分，或者是人工 cDNA 的结果（由于 cDNA 文库建立方法的原因，所以一些表面上未切割的 cDNA 实际上是由于在基因组 DNA 中以寡脱氧核糖腺核苷尾启动互补链的人工产物）时，会出现过度推测。推测不足是由于寻找具有非常小的外显子和表达有限的基因，以及寻找和确定编码非翻译 RNA 基因的困难所造成的。

8.3.6　人类基因组计划的最后阶段：基因注释和基因本体论

随着人类基因组测序到达最后阶段，人们致力于开发可以系统和方便地查阅到大量的人类基因组信息的新软件。**基因组浏览器**（genome browser）的精巧设计如 Ensembl（由 Wellcome Trust Sanger 研究所和欧洲生物信息研究所共同开发）提供了一个图表界面来描述各个染色体和染色体亚区的基因组信息。用户可以快速地把选定的人类染色体序列从宏观调整至核苷酸级，鉴别基因和相关的外显子、RNA 及蛋白质。调整的关

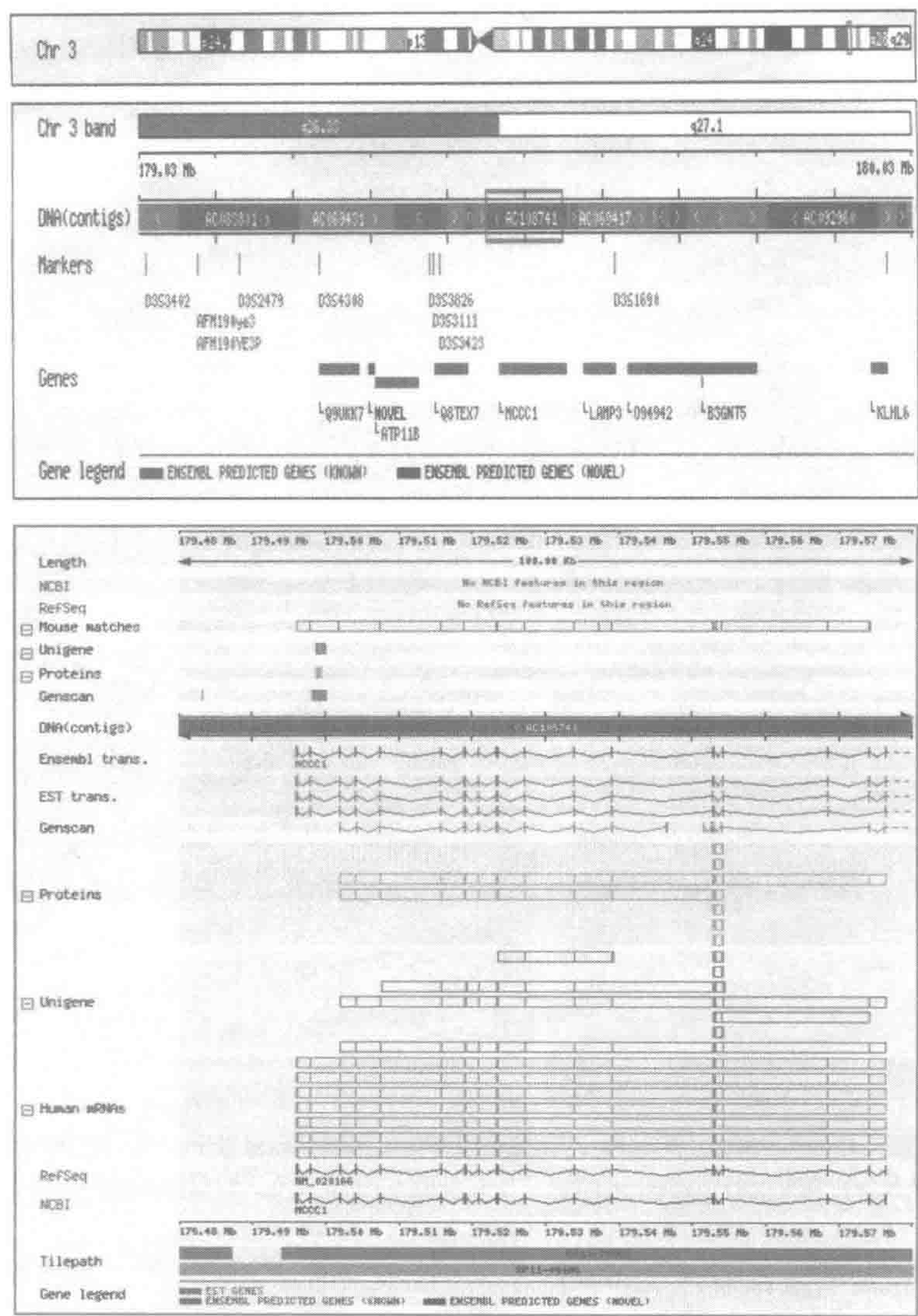


图 8.6　Ensembl 基因组浏览器的图形界面使得在各个染色体中查找组装的基因组序列十分方便

Ensembl 基因组浏览器（http：//www.ensembl.org）是 Wellcome Trust Sanger 研究所和欧洲生物信息研究所合作开发的，可以安染色体浏览查找人类和小鼠许多组装的基因组序列。最初选定人类 3 号染色体，再在染色体模式图上点击选定跨越 3q26.33-3q27.1 的染色体亚区（以上图的红色开放框表示）。选中的区域在中图中以扩大的形式表示，确定为从 179.03Mb 到 180.03Mb 的 1-Mb 结构。中图中的红色开放框是 100kb 的片段（179.48Mb 到 179.58Mb），在下图中以扩展窗口表示，阐明基因、外显子、RNA、蛋白质结构等。

键在于点击一结束工具，它保证了无数的与其他数据库和程序的连接（包括与其他组装基因组序列特别是小鼠序列的直接连接），并让用户追踪需要的信息（见 http：//www.ensembl.org/ 和图 8.6）。随着关于基因和其他功能单位的信息越来越多，在不断的周期性升级的基因组浏览器内可以得到更多的提供信息的和精确的**基因注释**（gene annotation）。

另一项主要的成就就是**基因本体论**（gene ontology），它建立了用来确定基因功能的系统和分类词汇表。**基因本体论协作组**［Gene Ontology（GO）Consortium］参与研究人类和其他基因组，并开发了一个确定能够应用于所有生物的基因功能的通用系统（the Gene Ontology Consortium，2000；2001；也可见http：//www.geneontology.org/）。通用词汇表可以用来在多个蛋白质和物种间查找共同点。基因本体论协作组开发了三个不同的本体论——生物进程，细胞成分及分子功能来描述基因产物，这些就可以解释各物种间分子特性。每个词汇表都被构建成定向的非环形图，其中任何一个术语都有一个以上的双亲和零个、一个或更多的孩子。这使其对生物学的描述比分类图要详尽得多。现在 GO 词汇表中有超过 11000 个术语，对其用途都做了及时的严格规定。

8.3.7　分析人类基因组序列变异对人类学和医学研究非常重要

人类基因组计划在其启动时的构思是获得相当于一个或者少数单倍体基因组的克隆化人类 DNA 片段集合的核苷酸序列。它并没有考虑到人类的遗传多样性。随后关于人类遗传多样性的信息在不同领域中受到重视：

- **人类学**。此信息应该有助于追踪人类起源、史前群体迁移和社会结构的人类学和历史学研究；
- **法医分析**。用于鉴定个人身份或证实亲密生物学关系（DNA 图谱）的 DNA 检测，其准确性部分依赖从一个人群到另一个人群诊断性 DNA 标记变化的理解；
- **医学研究**。一般疾病都是多因素的，鉴定潜在的致病基因是相当困难的（见 15 章）。近期人们致力于鉴定遗传的变异，它能解释不同人群对特定疾病的易感性增高或对某些疾病具有相对抵抗性方面的差异。

人类基因组多样性计划

Cavalli-Sforza 等（1991）提出了全球致力于人类基因组序列多样性研究的想法，即**人类基因组多样性计划**（Human Genome Diversity Project，HGDP）。该计划的重点优先集中在需要从大量的种群收集 DNA 样品。然而，虽然由 HUGO 支持，但是该计划仍有很大的困难（Greely，2001）。从一开始就陷入赞助资金的显著缺乏。只要该计划的基本目标单纯是寻找种群的标记，追踪人类迁徙的源头和祖先的血统，寻找大量资金赞助就不具有说服力。

HDGP 还被许多争论所困扰。在一些情况，研究者们走访位于生存边缘的隔离人群，迅速取得样本而不花费时间解释其意义，以后也很少或根本不进行交流。支持者们认为应该保护我们的文化遗产，而反对者则用诸如“直升机遗传学”这样的词汇来批评他们通常用麻木的“快进快出”的取样方法获得样本。

单链核苷酸多态性（SNP）图

虽然高度多态的微卫星标记遍布于整个基因组中，但是它们并不特别经得起自动化分型，并且大约每 30kb 左右才出现一次。**单链核苷酸多态性**（SNP）不是很多态的（典型的只有两个等位基因），但它们很容易自动分型，而且在基因组中很常见，平均约每 1kb 中出现一次（它们是如何分型的见节 7.1.3 和框 18.2）。故人类 SNP 图在构建

更高分辨率的遗传图中有重要应用，该遗传图是对预期含有常见疾病基因的染色体区进行研究所必需的。

国际 SNP 协作有限公司于 1999 年成立，是 Wellcom Trust 和大约十二个私人医药公司以非营利性质合作组成的。2 年内，它递交了一份约有 1.42 亿个 SNP 的人类 SNP 图，相当于每 2kb 就有一个 SNP（国际 SNP 图工作组，2001）。现在人们正致力于使用高密度 SNP 图定位常见疾病的致病基因，而更多商业兴趣也集中在药物基因组学的应用方面。

8.3.8 没有合适的安全保障，人类基因组计划可能导致对致病基因携带者的歧视以及优生学说的复兴

任何重要的科学的进展都会带来开发利用的恐惧。HGP 也不例外，计划有设想的好处，也有不利的一面。当我们掌握了人类的全部基因并可以检测大量疾病相关的突变时，对那些显示为携带致病基因的个体进行疾病靶向预防有很大的好处。但是，同样的信息也会用于对这些个体的歧视。比如，现在普遍的担忧是，保险公司可能会坚持让人们做大规模遗传筛查，以确定是否携带如糖尿病、心血管疾病、癌症以及各种精神疾病等常见疾病易感性的基因。然后那些碰巧鉴定为携带这样的疾病相关等位基因的非常健康的个体会被拒绝人身或医疗保险。很明显现在就有一小部分人受到这样的歧视；而值得注意的是这种歧视将来也许会施加到我们社会中相当大的一部分人身上。保证人们的不知情权也很重要。在所有遗传咨询和检测中一个基本的道德准则是只有在一个完全知情的成年患者明确要求下，才能给出相关的遗传信息。

另一个麻烦的地方就是生物学决定论的问题，以及对人类基因知识的广泛了解是否会促进优生学说的复兴即选择性孕育或者其他一些用来“改善”人类质量的遗传技术的应用（Garver and Garver，1994）。过去，一些国家（包括美国和德国）的负优生学运动歧视那些在某种程度上被认为是低等的个体，甚至强迫他们绝育。也存在利用**遗传增强**（genetic enhancement）专注于积极选择那些判定为合乎需要的遗传性状的可能性（21 章伦理框 3）。承认上述问题，美国人类基因组计划投入相当大的资源来支持伦理、法律和社会影响研究计划（例如 http://www.nhgri.nih.gov/ELSI）。

8.4 模式生物的基因组计划

人类基因组作图并不是人类基因组计划唯一的科学重点，一开始人们就清楚地认识到五种关键模式生物基因组测序的价值，而从那时起，这个目录就变成一个非常充实的目录。它包括多种单细胞微生物和各种多细胞模式生物，其中许多尤为适用于遗传分析。在某种程度上，小型基因组测序也被认为是大规模人类基因组测序。至 2003 年 5 月，140 多种有机体的基因组完成测序（或近似这样；见 http://wit.integratedgenomics.com/GOLD/；http://www2.ebi.ac.uk/genomes/）。

8.4.1 原核生物基因组计划具有高度多样性

原核生物基因组测序计划的多样性

许多原核生物模型已经建立很久了（框 8.7）。原核生物基因组一般比较小（通常只有一个或几个 Mb），它们特别适于相对快速的测序，导致很多**原核生物基因组计划**（prokaryotic genome project）的迅速发展。至 2003 年 5 月，已经完成了总计 122 种原核生物（16 种古细菌和 106 种细菌）基因组的测序，另外还有 342 种（23 种古细菌，319 种细菌）的基因组测序正在进行中（Doolittle，2002）。

框 8.7　单细胞生物模型

各种单细胞生物非常适用于遗传和生化分析，具有传代速度很快、容易大量培养等重要优势。虽然它们在进化历程上离我们非常远，但它们对各种科学和医学研究还是很有用处的。可以在各个广阔的研究领域给遗传学家和医学研究人员带给新的思路，有些具有明确的医学应用，包括：

- 基因功能和细胞进化：各种关键的重要核心细胞活动在进化过程中非常保守。通过研究单细胞生物同等的基因和进程，我们可以了解哺乳动物的基因功能和关键性细胞活动的性质，如核糖体生物合成、细胞周期调控、膜转运、聚合酶功能等；
- 发病机制：已知许多单细胞生物可以导致疾病，多年来医学都极力寻找合适的有效药物或其他治疗方法。通过确定这些致病生物的全部基因组，以及研究疾病发生的确切分子基础，我们可以期待在如何应对这些致病性微生物方面有新的思路；
- 进化：序列分析提供了生物如何相互关联的最有用的方法。因此，分析 rRNA 序列可以突破性的把生物分成三个基本部分：古细菌，细菌，真核生物（框 12.4）；而基因组序列的完全比较无疑可以带给我们更多重要的思路（图 12.22）。

细菌（BACTERIA）

各种正常情况下非致病性的细菌长期以来就是很好的模式生物，尤其是存活在人类和其他脊椎动物消化道中的棒状的**大肠杆菌**（以共生状态存在：其贡献在于合成我们需要的维生素 K 和维生素 B 复合物）。几十年的深入研究使得我们对大肠杆菌的了解比其他任意一种细胞都多，我们对于生命基本机理的理解——如 DNA 复制、转录、蛋白质合成等——大部分都来自于对这种生物的研究。

当然还有各种致病菌可以造成不同严重程度的疾病。其中也包括一些大肠杆菌，它们可以引发脑膜炎、败血症、泌尿系感染以及肠道感染。一个重要的例子是大肠杆菌 0157：H7，作为过去获得一种特定的噬菌体序列的结果，它就会发生基因修饰而可以产生一种毒素。在一些病例中，该毒素会导致儿童或成人发生致命的大量出血；在其他病例中，会导致肾衰。人们已经启动了各种基因组计划试图获得致病性微生物的全部序列（表 8.3）。

古细菌（ARCHAEA）

古细菌是原核生物。表面形态类似细菌，但它们在进化非常早期就偏离细菌了。最初在一些稀有的、通常非常极端的环境中发现它们——在温泉或深海裂隙附近样的高温环境中、在极端 pH 或盐度的水中、在缺氧的沼泽泥土中、在地下深层沉积的石油中及在海洋底部。但现在，人们知道它们存在于一些我们更加熟悉的地方如土壤和湖泊，并且在牛、白蚁、海洋生物的消化道中茁壮成长，制造沼气。

框 8.7　单细胞生物模型（续）

古细菌的代谢和能量转化系统和细菌的很相似，但处理和加工遗传信息（DNA 复制、转录、翻译）的系统却与真核生物更为接近而不是细菌。现在还没发现古细菌与疾病有关，而研究其基因组主要兴趣是它们作为一个与其他生命形式截然不同的界面存在，以及想更多了解为什么它们进化得如此独特。

酵母（YEAST，单细胞真菌，UNICELLULAR FUNGI）

酵母是一种单细胞真菌，所以是真核生物。它们通常存在于植物的叶与花朵、土壤和海水中，也共生或寄生在恒温动物的皮肤上或消化道内。它们通常以芽殖而不是二分裂进行复制：在一新的细胞壁形成而分开为二之前，来自亲本细胞的细胞质和分裂的核在初期是一个具有芽或子酵母的统一体。

酵母是一种很有价值的模式生物，因为已知许多关键的分子从酵母到哺乳动物都是高度保守的。更重要的是，人们发现许多酵母细胞周期和 DNA 修复基因所对应的人类基因直接参与人类细胞分裂。功能异常会导致癌症或先天缺陷。在酵母细胞中研究这些基因的正常功能非常容易，在实验室中操作也很方便。这可以帮助我们了解这些基因对人类的重要作用，并有助于在更复杂的细胞类型中指导实验。因此研究酵母细胞分裂的调控是与人类健康和对很多临床疾病的了解紧密联系的。一般来说酵母不与疾病相关，但有些致病性酵母——特别是念珠菌属——会导致一种常见的健康问题。

- **酿酒酵母**（*saccharomyces cerevisiae*）是一种出芽酵母，长期以来在发酵和酿造上有很重要的作用。因其简单、容易生长，所以很适合基础研究，是研究最广泛的真核生物之一。部分由于其高频率的非同源重组，所以很适合进行遗传研究。它一直被用作细胞生物各方面分析的模型，如细胞周期调控、蛋白质运输和转录调节。
- **粟酒裂殖酵母**（*schizosaccharomyces pombe*）是一种二分裂酵母，世代时间很短（2～4h）。其最近才被广泛研究，主要作为细胞周期调控（周期中有一个特殊的 G2 期）和分化的模型。它与酿酒酵母菌的关系很远，但在染色体结构和 RNA 加工的某些方面，它更接近于高等真核生物而不是酿酒酵母。
- **白色念珠菌**（*candida albicans*）自然存在于健康人的口腔中，但可以导致炎症，有时如果个体的免疫功能异常还会导致严重的感染，如阴道和口腔溃烂、尿布湿疹等。

原生动物 PROTOZOA（单细胞动物）

原虫是一大类单细胞动物（无光合作用），包括**阿米巴**（*amebae*）、**鞭毛虫**（*flagellates*）和**纤毛虫**（*ciliates*）（各自借助伪足、鞭毛、纤毛运动），以及其他具有复杂生命周期的生物。生物医学家对它们很感兴趣，因为它们可以作为细胞和发育生物学各种方面的模型，而且许多原虫还是致病性寄生虫。后者可以是致病性细菌的宿主（引起疾病如军团病、沙门菌病、结核等），或直接导致一些疾病，如：

- **痢疾阿米巴**（*Entamoeba histolytica*）：一种导致严重消化道疾病的寄生性阿米巴；
- **锥虫**（*Trypanosomes*）：一种导致热带昏睡病的寄生性鞭毛虫；
- **贾第虫**（*Giardia*）：一种导致严重腹泻的寄生性鞭毛虫；
- **疟原虫**（*Plasmodium*）：导致疟疾。
- **弓浆虫**（*Toxoplasma*）：导致消化道疾病和内脏损伤。

作为细胞和发育生物学的模型，人们最感兴趣的是**盘基网柄菌**（*Dictyostelium discoideum*），一种所谓的群居性阿米巴：虽然平时各个细胞独自生长，但在面临一些恶劣环境如饥饿时，细胞就会相互作用组成多细胞生物。近 100000 个细胞通过释放化学引诱物 cAMP 相互传递信号，并通过

框 8.7　单细胞生物模型（续）

趋化作用聚集成一团，周围有细胞外基质包绕。这种形成多细胞生物体的方法与多细胞动物胚胎形成的早期步骤是完全不同的。但是在网柄菌和多细胞动物中随后的过程都依赖于细胞－细胞交流。这种生物非常独特地适合进行胞质分裂、运动性、胞吞作用、趋化性、信号传导的研究，以及一些发育方面的研究如细胞分类、模式形成、细胞类型决定。许多这些地细胞行为和生化机制在其他模式生物要么缺乏，要么无法利用。

纤毛虫不像其他原虫那样被广泛研究。其中研究最多的就是**四膜虫**（*Tetrahymena*），它是一种淡水生物，常见于河流、湖泊和池塘，而**嗜热四膜虫**（*Tetrahymena thermophila*）的基因组计划正在积极筹备中（Turkewitz *et al*.,2002）。四膜虫的细胞很大（沿前后轴有 40～50μm），和其他纤毛虫一样，它有许多非常复杂而独特的细胞结构。作为典型的纤毛虫，四膜虫的细胞核由结构和功能不同的两部分组成，称之为核的二态性。小核是生殖系，即储存有性繁殖的遗传信息。小核是二倍体，有五对染色体。大核（MAC）是体细胞核，即在旺盛生长过程中活跃表达的核，MAC DNA 并不遗传用于有性繁殖。四膜虫是一种完善的细胞和发育生物学模型，主要用于研究圆形细胞运动、发育过程中程序性 DNA 重排、调节性分泌、胞吞作用和端粒维持和功能。

第一个完成（1995）的原核生物基因组是 1.83Mb 的流感嗜血杆菌（*Haemophilus influenzae*）基因组。这是一个里程碑：首次完成了对非寄生的有机体基因组的测序（Tang *et al*.,1997）。随后，又实现了许多其他的之最：最小的自主性复制实体基因组（生殖器支原体，1995），第一个原虫基因组（詹氏甲烷球菌，1996），以及随后获得 4.6Mb 大肠杆菌基因组（*E. coli* genome）全部序列的重要成就（Pennisi，1997）。

已完成基因组测序的原核生物的列表揭示了不同的优越性。在一些例子中，驱动力是为了解不同有机体之间的进化关系，如对于古细菌基因组来说（Olsen and Woese，1997），而对于生殖器支原体来说，它是为了明确最小基因组的组成，这是已知的最小的细胞基因组（现已知仅有 470 个基因）。在另外一些例子中，如在大肠杆菌和枯草杆菌中，其优势仅仅是推动常用实验有机体的基础研究。对于许多研究者来说，最有价值的就是大肠杆菌——研究最集中的细菌。但令人惊奇的是，尽管做了大量的前期研究，起始鉴定的 4288 个基因中约 40%仍没有已知的功能，进而成为深入研究的主题。然而，对许多其他生物来说，基因组测序的主要动机是它们的医学相关性。

疾病相关的原核生物基因组计划

在某些情况下，原核生物由于其已知的与某些慢性疾病的相关性（Danesh *et al*.，1997），或者由于它们是疾病致病因素（表 8.3）而被选择进行基因组测序。除了对这些生物有了进一步全面的了解之外，这些新信息可用于开发更为敏感的诊断手段及药物/疫苗研发的新靶点。

表 8.3　微生物战争：致病性微生物基因组计划的实例（进一步阅读见互联网资源）

有机体	基因组大小(染色体数量)	相关疾病
细菌		
炭疽杆菌	4.5Mb(1)	炭疽
百日咳杆菌	3.88Mb(1)	百日咳
博氏疏螺旋体	0.95Mb(1)	Lime 病
肺炎衣原体	1.0Mb(1)	呼吸系统疾病;冠心病
沙眼衣原体	1.7 Mb(1)	沙眼,一种致盲原因
难辨梭状芽孢杆菌	4.4 Mb(1)	抗生素相关腹泻;伪膜性肠炎
幽门螺旋杆菌	1.67 Mb(1)	消化性溃疡
麻风分枝杆菌	2.8 Mb(1)	麻风病
结核分支杆菌	4.4 Mb(1)	结核
普氏立克次体	1.1 Mb(1)	斑疹伤寒
伤寒沙门氏菌	4.5 Mb(1)	伤寒
梅毒螺旋体	1.1 Mb(1)	梅毒
霍乱弧菌	2.5 Mb(1)	霍乱
鼠疫耶氏菌	4.38 Mb(1)	鼠疫
原生动物		
硕大利什曼原虫	33.6Mb(36)	Leishmaniasis 病
镰型疟原虫	23Mb(14)	疟疾
布氏锥虫	54Mb(22)[a]	非洲锥虫病(昏睡病)
克氏锥虫	87Mb(>42)	美洲锥虫病(Chagas 病)

a 组成 11 对。

8.4.2　酿酒酵母基因组计划是许多个成功的原生生物基因组计划中的第一个

原生生物（protist）是单细胞真核生物，包括单细胞动物（原生动物），单细胞真菌，以及单细胞植物。人们已经进行了许多原生生物基因组计划，有些时候为了了解基本模式生物的需要，而另些时候是作为对抗致病性原虫所引起的疾病的手段。

酿酒酵母基因组计划

酿酒酵母（*saccharomyces cerevisiae*）是一种单细胞真菌，长期以来一直是一种很好的真核模式生物，部分是由于其容易进行遗传分析（框 8.7）。欧美协会对其 16 条染色体进行了测序，而 Goffeau 等报道了它的全部序列（1996）。这是生物学上的又一个里程碑：首个真核细胞的全部序列。数据显示酵母基因紧密成簇排列，平均间隔 2kb。在 6340 个基因中，约 7％为非翻译 RNA。虽然它是研究最多的生物之一，但首次报道其序列时，仍有 60％的基因没有实验性确定的功能。然而，酵母基因相当大的一部分与哺乳动物有明确的同源性，所以只有大约 25％的酵母基因无论如何也没有其功能的线索（Botstein *et al.*,1997）。总之，这项计划的成功完结现已开启了大规模功能分析

(19 章)。

粟酒裂殖酵母基因组计划

裂殖酵母菌是另一种单细胞真菌，也是一直受人喜爱的模式生物（框 8.7）。Wood 等（2002）公布了全部 13.8Mb 的序列，发现了总计为 4824 个蛋白质编码基因。数据显示酿酒酵母菌和粟酒裂殖酵母菌基因之间有相当大的不同，有几百个基因在粟酒裂殖酵母菌中存在而在酿酒酵母菌中明显缺失，相反，在内含子数量（粟酒裂殖酵母菌有 4700 个而酿酒酵母菌仅有 275 个）和转座元件（与酿酒酵母菌相比，粟酒裂殖酵母菌中很少）等方面也有差别。

镰状疟原虫基因组计划

Gardner 等（2002）对镰状疟原虫基因组测序的报道是另一个里程碑：首次测序了真核生物寄生虫基因组。镰状疟原虫是一种致命性的疟疾寄生虫，与此同时关于其宿主冈比亚按蚊（*Anopheles gambiae*）（节 8.4.4）基因组全部序列的报道为治疗疟疾开辟了新的天地，疟疾是一种常见的致死性疾病，每年有一百万人死于这个疾病，大部分都发生在亚撒哈拉沙漠的非洲。

其他原生生物基因组计划

其他原生生物基因组计划包括其他各种参与人类寄生虫感染的**致病性原生动物**（pathogenic protozoa）的基因组计划。在大多数情况下，基因组大小都是实在的，一般在 30～90Mb 范围内（表 8.3）。另外，人们还进行了其他已充分研究过并容易进行生化/遗传分析的生物的基因组计划，包括**构巢霉菌**（*Aspergillus nidulans*）和**粗糙链孢霉菌**（*Neurospora crassa*）。

8.4.3　秀丽新小杆线虫基因组计划是首个完成的动物基因组计划

秀丽新小杆线虫基因组计划

虽然秀丽新小杆线虫是一种简单的生物，只有 1mm 长，但它却被认为是发育的重要模型，对模拟与人类细胞相关的其他过程也是十分有用的（框 8.8）。由于秀丽新小杆线虫的基因组较大（近 100Mb），因此它的基因组计划也被认为是人类基因组大规模测序的主要先导模型。该基因组计划成功地由 Wellcome Trust Sanger 研究院和华盛顿大学医学院完成（秀丽新小杆线虫测序协作组，1999）。这是另一项里程碑式的成就，第一次提供了多细胞动物（后生动物）的遗传指令。

秀丽新小杆线虫基因组计划最初报道了共有近 19000 个编码多肽的基因和 1000 多个编码非翻译 RNA 分子的基因，基因平均间隔 5kb。惊人数目的基因看起来好像作为**操纵子**（operon）的一部分出现，在操纵子内，每个基因转录为大的多基因 RNA 转录物的一部分。在最初报道的同时，将其与其他各地公布的序列信息比较后发现，新近鉴定的秀丽新小杆线虫基因中约三分之一与先前已知的基因类似，12000 多个编码多肽的基因功能未知，大多数是推测的基因而没有进行实验证实。

对于报道的 19000 个编码多肽的基因中，只有 9000 多个基因有实验数据支持，而剩下近 10000 个预测的基因都仅是通过基于计算机的序列分析而鉴定出来的。验证这些基因相应的 RNA 转录物的后继分析表明，至少 80%计算机推测的基因是正确的，由此 Reboul 等（2001）推断秀丽新小杆线虫至少有 17300 个基因。现在大量工作致力于特定基因功能的研究以及通过大规模化学诱变制造大量的突变表型。

框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型

广范围的多细胞动物模型已用于了解发育和细胞生物学的基本过程，了解基因功能，以及用作疾病模型。它们从无脊椎动物的蠕虫和苍蝇到各种鱼类、蛙类、鸟类和哺乳动物变化。见 Hedges（2002）及图 12.23 和 12.24 的系统发生图。

秀丽新小杆线虫（CAENORHABDITIS ELEGANS）（蛔虫）



秀丽新小杆线虫是一种线虫或蛔虫（与扁虫和节虫不同）。蛔虫在地球上的数量远远多于其他复杂的生物，在气候温和的环境中到处可见，尤其是在土壤里。它们可以是非寄生的（如秀丽新小杆线虫）或寄生生存的。约十亿人感染蛔虫，蛔虫传播河盲、象皮病，并吞吃农作物。

秀丽新小杆线虫有 1mm 长。它们有两种性别：雄性（XO）；雌雄同体（XX），即一种改变的雌性。雌雄同体是主要性别。通过精子和卵子它可以进行自我受精，形成等位基因的纯合性。雌雄同体的两个 X 染色体偶然缺失一个就会形成雄性，雌雄同体在可能的情况下会优先与雄性交配。细胞谱系是固定的，恰好成年雌雄同体有 959 个体细胞，而成年雄性有 1031 个体细胞。

秀丽新小杆线虫是一种重要的发育模型，实验室内培养容易（在琼脂平板上以细菌喂养，或在液体培养基中培养），并且易于进行遗传分析。对遗传学分析来说，除了标准的 DNA 水平的敲除基因功能方法以外，**RNA 干扰技术**［RNA interference（RNAi）technology］可以使特定基因的表达瞬时失活。在这种方法中将感兴趣的基因的双链 RNA 注入到卵母细胞中。该双链 RNA 使同源基因的表达失活，产生的突变表型可用于研究寻找基因功能改变的线索（节 20.2.6）。

秀丽新小杆线虫的一些特点使其成为发育生物学和相关研究的一个良好的模式生物：

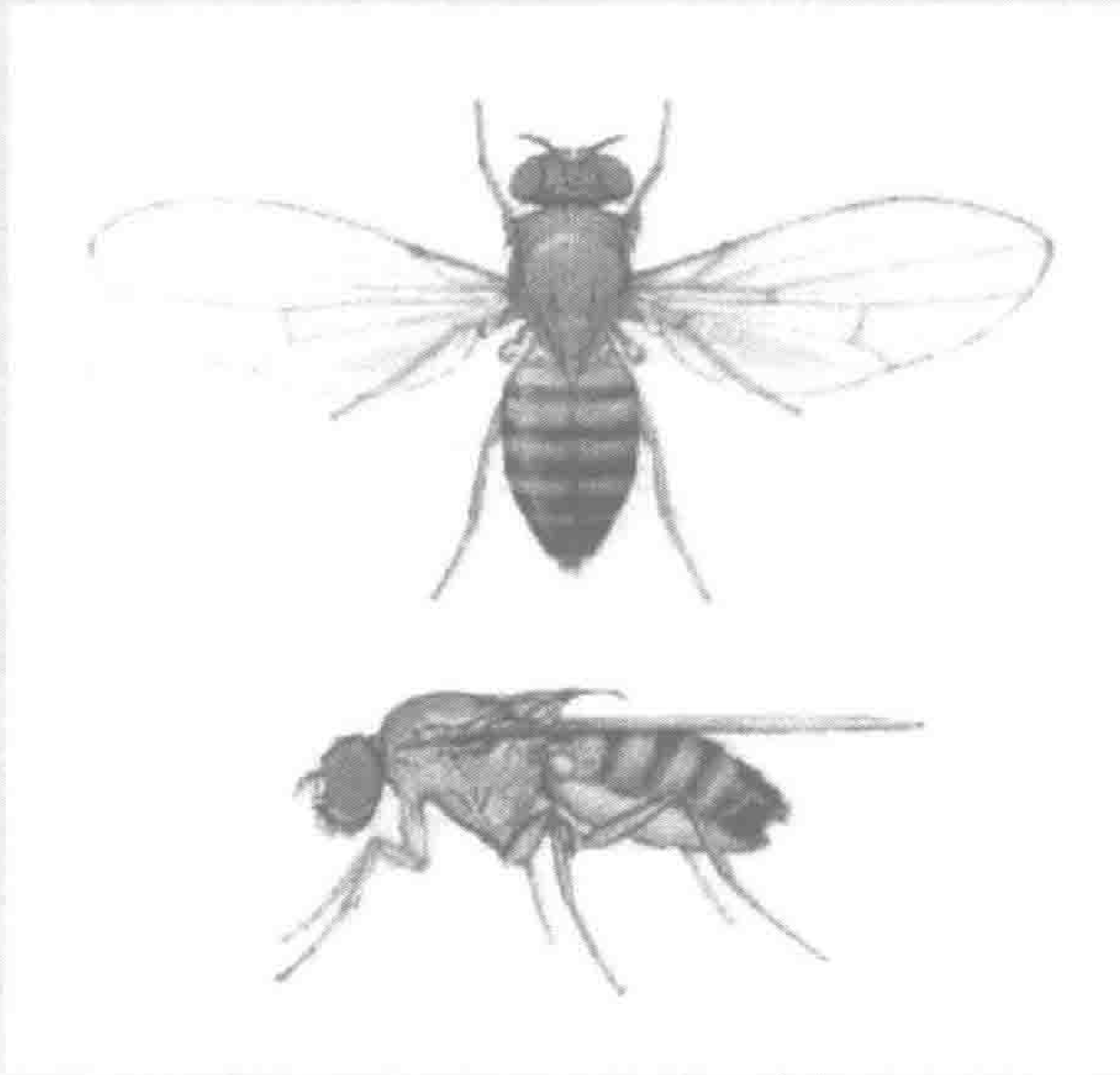
- 表达研究。因为秀丽新小杆线虫在生命过程中处于透明状态，所以可以把绿色荧光蛋白基因（GFP，框 20.4）与其基因连接起来，用于寻找基因在线虫的何处表达；
- 谱系研究。秀丽新小杆线虫的透明性意味着每个细胞在发育过程中都可以观察和追踪。结果，由于细胞谱系的固定性，我们就可以了解秀丽新小杆线虫每个细胞的确切谱系来源——这个信息在所有其他多细胞生物是无法了解的。
- 神经系统。它具有完整的神经系统线路图：全部 302 个神经元及其神经连接是已知的。秀丽新小杆线虫也拥有大多数脊椎动物脑部已知分子组成的基因。尽管许多科学家认为人脑非常复杂，我们绝不可能有希望完全掌握它，但了解秀丽新小杆线虫简单的神经系统可以提供许多思路，而了解所有秀丽新小杆线虫的基因对于了解它的神经系统很重要。

框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型（续）

▶ 衰老。这项研究很容易进行，因为线虫在三天之内从单细胞发育至完全个体，但只能生存 2 周。已经鉴定了许多突变，这些突变会导致秀丽新小杆线虫寿命延长，有些突变体生存时间会超出野生型线虫五倍以上。

▶ 凋亡。涉及**凋亡**（apoptosis）（程序性细胞死亡）的基因通常都是高度保守的，人们已经充分了解了秀丽新小杆线虫发育过程中的凋亡模式（雌雄同体最初有 1090 个细胞，但在发育过程中有 131 个细胞程序性死亡）。

黑腹果蝇（DROSOPHILA MELANOGASTER）



果蝇之所以叫这个名字是因为它们喜好吃水果。它的生命周期很短，特别适合进行复杂的遗传分析，几十年来人们对其进行了广泛的研究。现在人们对许多突变体进行了系统的研究，积累了大量基因功能的信息（见 Perrimon 于 1998 年关于基因功能研究新进展的摘要）。许多特性和方法都有助于基因作图和功能分析：

▶ 多线染色体（polytene chromosomes）。这些是间期染色体，长 2mm，存在于幼虫的唾液腺细胞中。其独特性在于它们是重复复制产生的，而不分到子细胞核中，形成正常单个 DNA 双链的 1024 个拷贝，如同盒中吸管一样并行排列在一起。由于这种平行的 DNA 扩增，所以在间期染色体中多线染色体是独一无二地能够在光镜下看到。其延伸的构象使得人们可以通过原位杂交确定染色体断裂点（上万个碱基）和 DNA 克隆位置；

▶ **P-元件**（P-element）。这个果蝇的转座元件使得人们可以进行几种实验室操作，包括诱发突变（Spradling *et al*.,1995）和转基因（节 20.2.2）。邻近的 P 因子插入片段之间的不等重组也可以产生精确的缺失；

▶ 利用条件基因表达系统 GAL4-UAS 有可能实现转基因在空间和时间上的限制性表达。可以进行大规模的诱发突变筛查，最近 RNAi 技术（见上文）可以用于瞬时失活特定基因。在许多情况下（大概 12000 个果蝇基因中的 2/3）功能缺失并不会产生突变表型，但转基因的错误表达常常通过产生显性/显性负相表型而提供基因功能的线索。对显性突变表型的抑制子/增强子进行一代筛查可以鉴定相互作用的基因。酵母 flp－frt 重组酶系统可以诱导有丝分裂克隆，因而形成纯合子片段，在发育晚期观察到致死性隐性突变表型。有丝分裂重组也可用于评定克隆中突变表型的一代筛查，以及发现影响晚期发育的致死性突变。

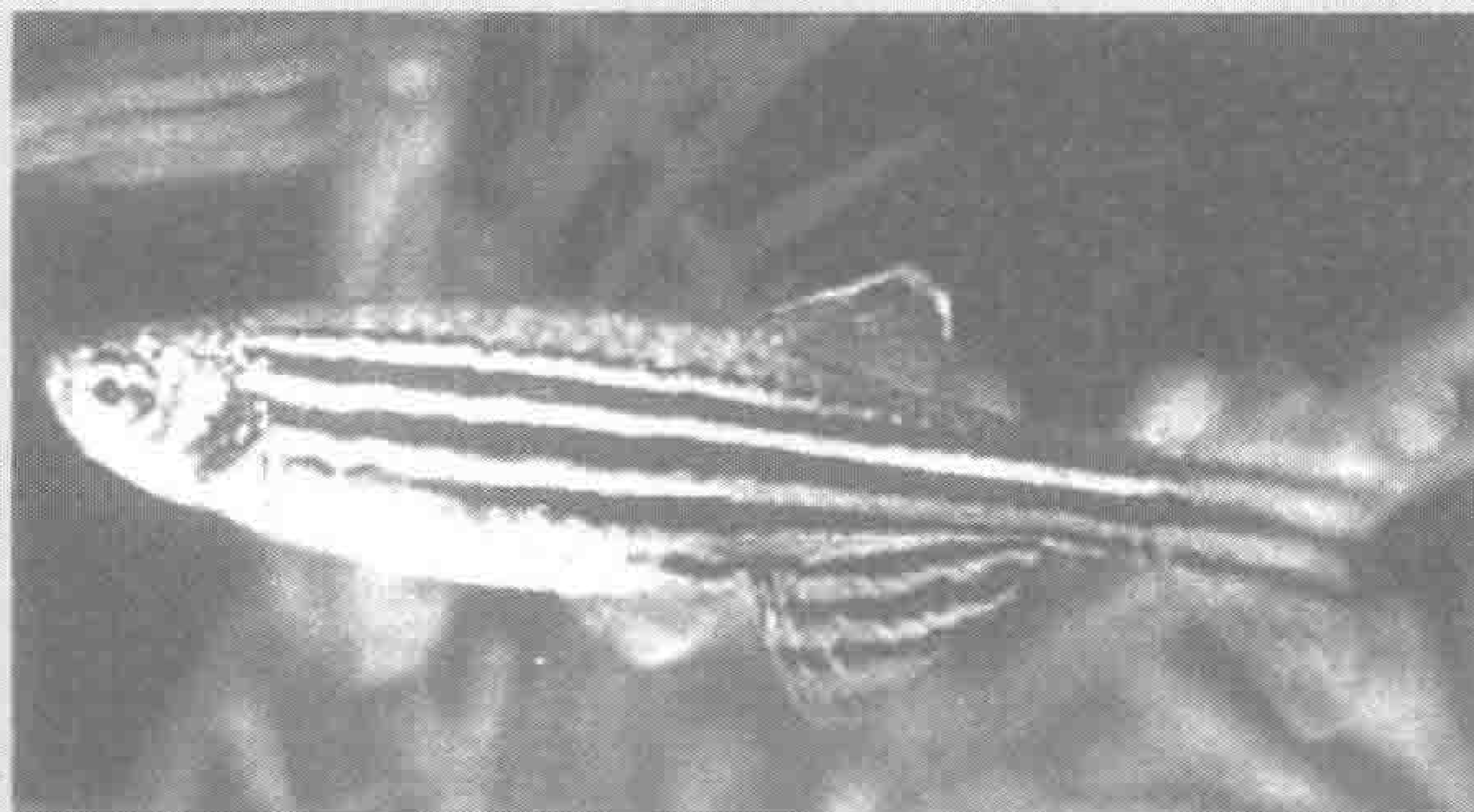
框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型（续）

虽然果蝇是无脊椎动物，其基因数目只有人类基因的一半，但人类和果蝇基因之间却有一些明显的相似之处。基因数量的不同是因为基因复制事件，在人类造成很大的基因家族；而大部分果蝇基因都具有人类基因的同源物，其中包括一些导致遗传疾病和癌症的基因。同源性的级别使人们可以成功地通过电子筛查鉴定出与果蝇突变基因相对应的人类 cDNA（Banfi *et al.*，1996）。许多高度保守基因在癌早期发育中具有重要作用，这些基因在果蝇中已经被很好地了解，而且早期发育和一些其他细胞过程中的一些相关途径从果蝇到哺乳动物基本上是保守的。因此，果蝇可以作为研究与人类系统直接相关的基因功能和基因相互作用的模型系统。

鱼类

不同的鱼类已用作模型，尤其是斑马鱼，它是一种很好的发育模型，河豚鱼因其非常紧凑的基因组，逐渐成为日益重要的疾病模型，而最近青鳉也被用作一种重要的发育模型。对这三种动物进行比较基因组研究，为脊椎动物基因组进化提供了重要的思路。

▶ **斑马鱼**（zebrafish）［斑马鱼（*Brachydanio rerio*）］



斑马鱼是起源于印度河流中的一种小型淡水鱼，现在常作为一种观赏鱼遍布世界。它的传代时间很短，并且每次交配可以产生大量的卵。它是脊椎动物发育的主要模型：在外部受精使人们可以观察到发育的全部过程，透明的胚胎促进了发育变异体的识别。脊椎动物发育过程中重要的基因通常是高度保守的，所以在斑马鱼很容易找到人类发育调控基因的类似物。大规模诱变筛查产生大量的有价值的发育变异体，其中有些可以用作模拟人类疾病（框 20.6）。RNA 干扰技术（见上文）也被用来进行特定基因的失活。

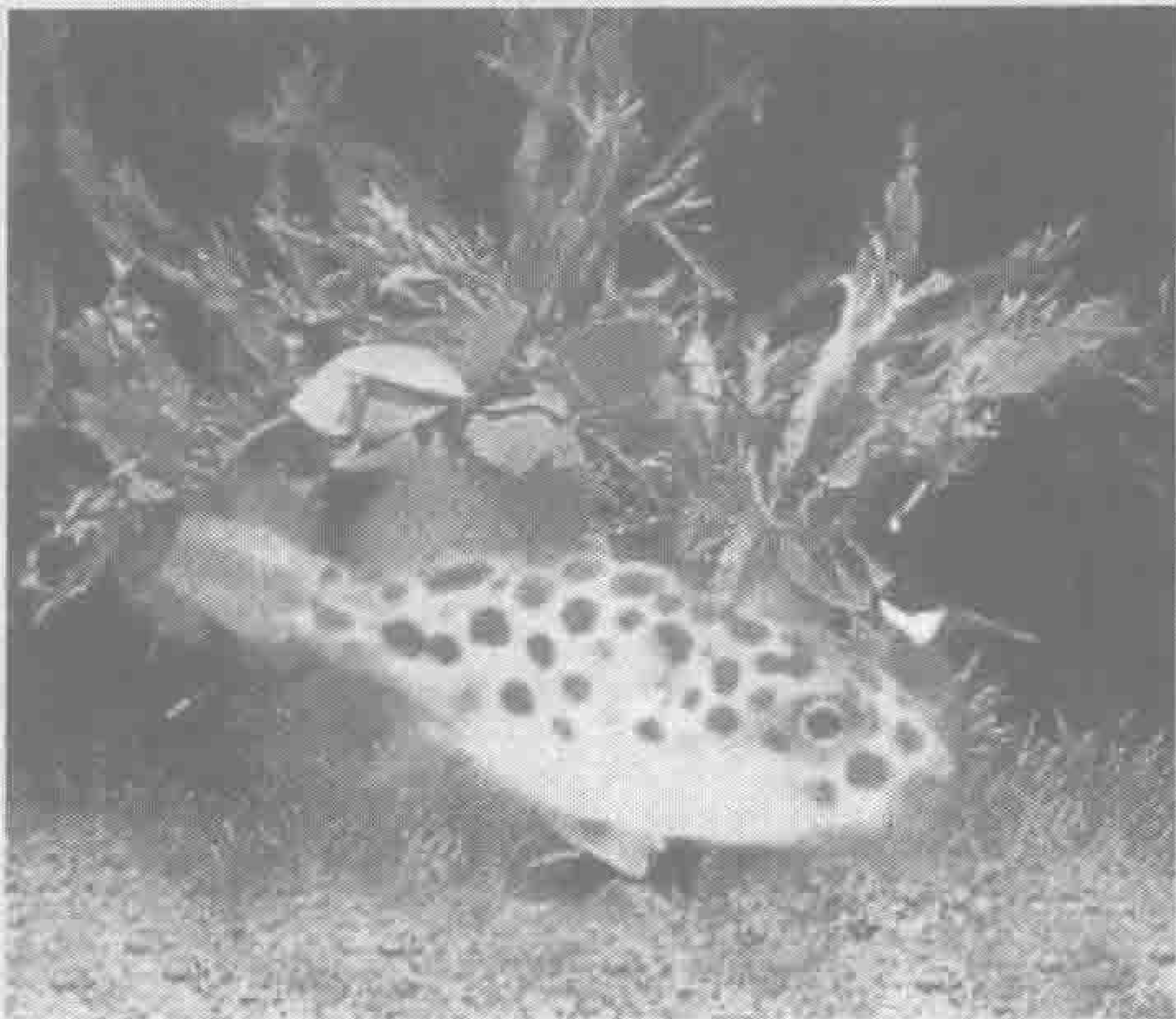
▶ **河豚鱼如红鳍多纪鲀**（*Takifugu rubripes rubripes*）（下面）和**黑青斑河豚**（*Tetraodon nigroviridis*）（下页）。



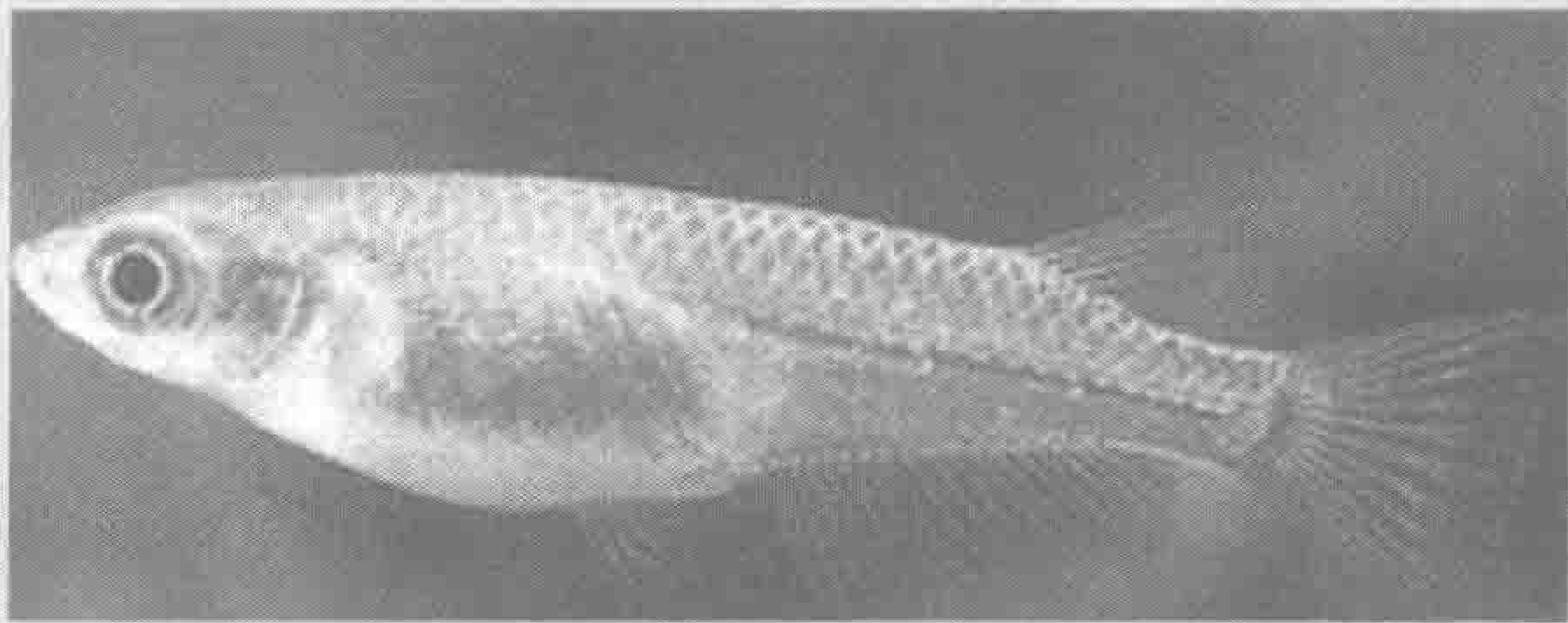
其应用价值在于比较基因组学（12.3 节）。河豚鱼的基因组非常密集，和哺乳动物相同数量的基因压缩在仅有人类或小鼠七分之一的基因组中。其外显子和重要调节序列都是保守的，可以通过比较作图来鉴定人类的同源基因（Clark，1999）。

框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型（续）

▶**青鳉**（medaka）



青鳉（*Oryzias latipes*）是一种小型卵生淡水鱼，主要分布在日本。它是斑马鱼的远亲（二者在 1.1 亿年前由同一祖先分化而来），也逐渐成为日益重要的发育模型。同斑马鱼一样，其传代时间很短（2～3 个月）、种系纯、有遗传图、转基因发生、增强子捕获、可以得到干细胞，因此适合进行遗传和胚胎学分析（Wittbrodt *et al*.,2002）。对青鳉和斑马鱼进行发育筛查发现的表型揭示二者的胚胎致死表型谱没有重叠。



鸡类



鸡类作为发育模型有几个优点。与哺乳动物一样，鸟类是羊膜动物（胚胎有羊膜），其发育与哺乳动物非常类似。但是，哺乳动物胚胎依靠母体获得营养（通过胎盘交换），而鸟类胚胎没有胎盘，是一种自主发育系统。因为鸡胚在体外发育，所以它可以在发育的所有阶段获得。除了获取容易之外，鸡胚还具有比较大，相对透明的优点，容易进行复杂的显微外科操作。因此，它提供了一个经典胚胎学与分子研究相结合的极佳系统（Brown *et al*.,2003）。

鸡胚常见的实验操作有：外科操作和组织移植；反转录病毒介导的基因转移；发育胚胎的电穿孔；胚胎培养。另外，鸡类提供了一个研究细胞进程的独特系统：**DT40 细胞系**（DT40 cell line）。鸡类 DT40 细胞可以不停地使其免疫球蛋白基因多样化；在可用的脊椎动物细胞系中，DT40 细胞是唯一可以以接近非常规重组的速度进行同源性重组的。结果就可以相对便利地在培养细胞系中进行靶基因的缺

框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型（续）

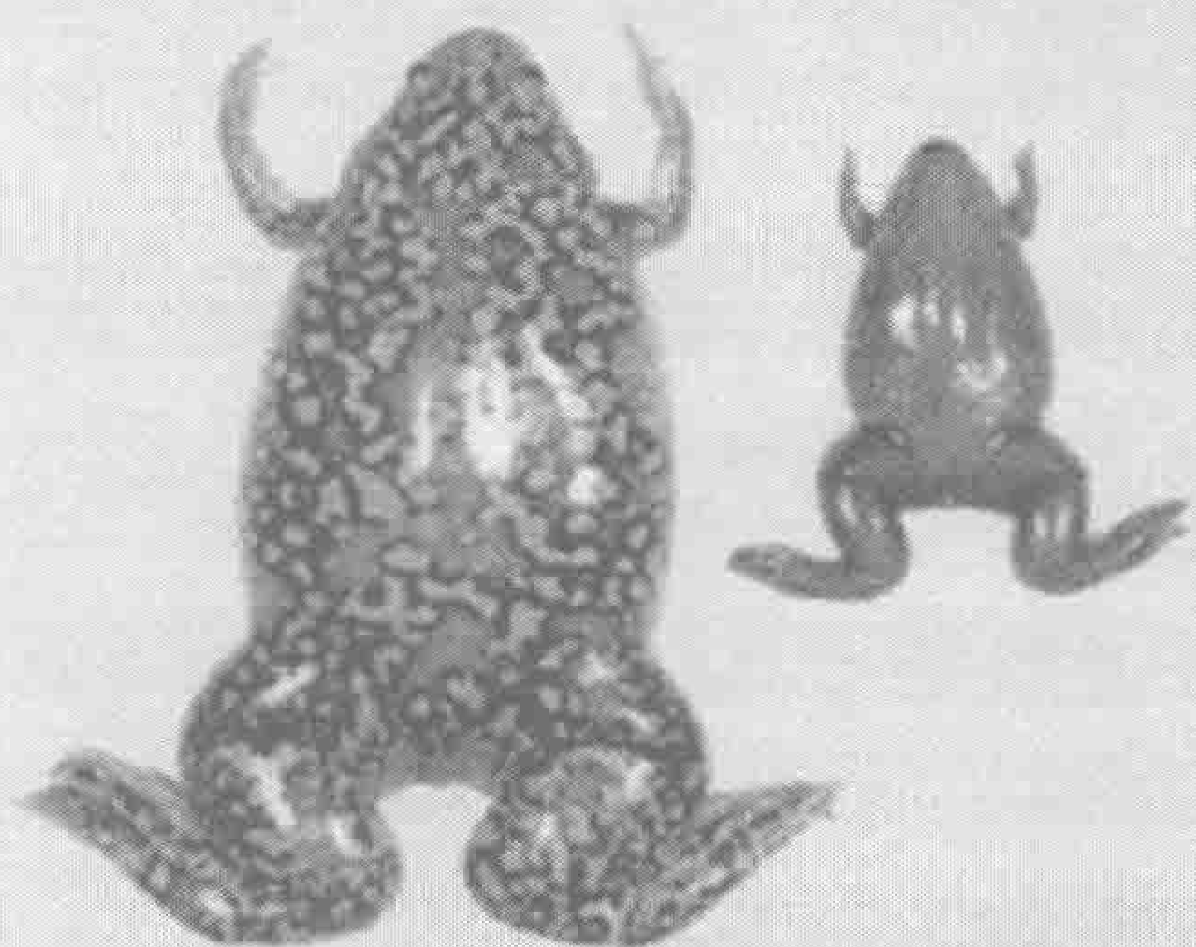
失和突变分析。鸡类转基因学，胚胎干细胞技术，精子、胚盘细胞、原始生殖细胞和胚胎干细胞的冰冻保存等技术快速发展。这些更新的技术将有助于建立一个以鸡类为基础的系统用于构建人类疾病模型。

爪蟾（非洲爪蛙）

这种非洲青蛙之所以这样称呼（Xenopus 即奇怪的足）是因为它强壮有蹼的后足有锋利的爪。长期以来它一直是胚胎发育和细胞生物学的良好模型：可以对其所有发育阶段进行观察，卵相对较大，胚胎早期就可以进行显微操作，如显微注射（mRNA、抗体和反义寡核苷酸）、细胞移植和标记实验。20 世纪 90 年代发展了有效的制作转基因胚胎的方法（Beck and Slack，2001），现在人们又在计划进行热带爪蟾诱发突变的筛查。

▶**光滑爪蟾**（*Xenopus laevis*）（左上图）

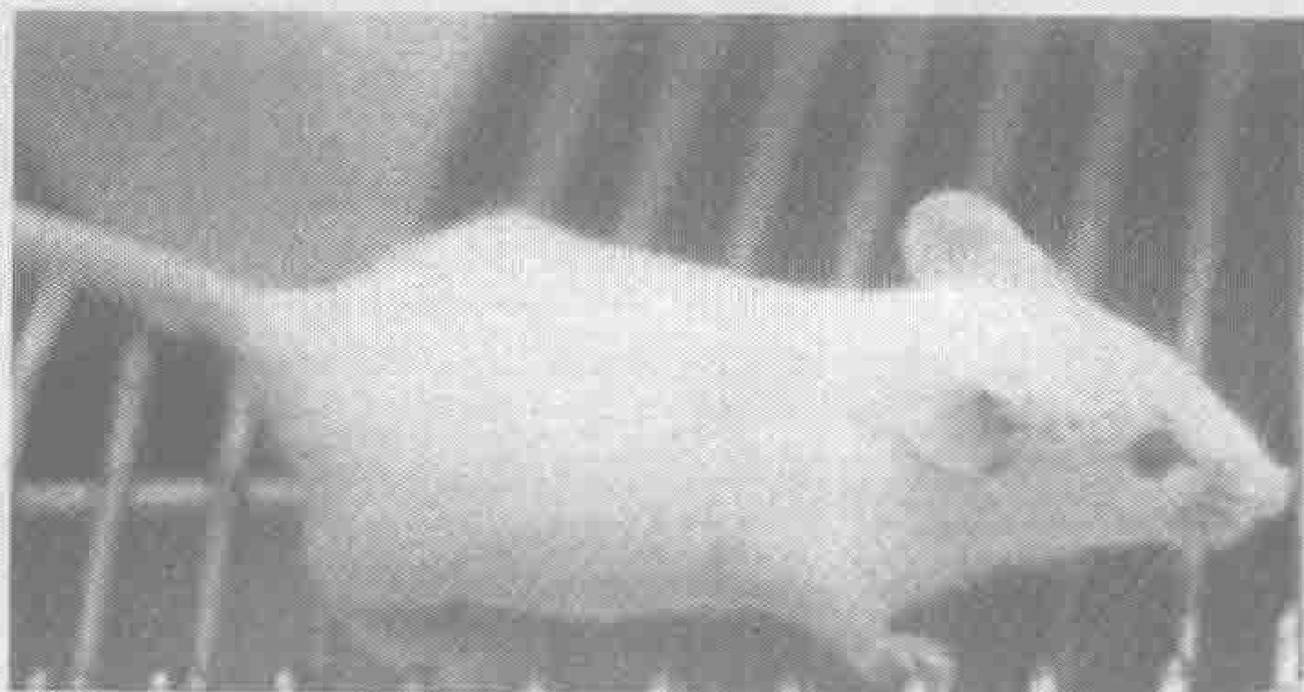
光滑爪蟾世代时间较长，为 1～2 年，每次可以产生 300～1000 枚大的（1～1.3mm）卵。它是研究早期命运决定、基本身体模式形成以及器官发生的重要模型。在细胞生物学和生物化学方面的贡献包括染色体复制、染色质和核的组装、细胞周期组分、细胞骨架元件及信号通路的创新性工作。光滑爪蟾的基因组是异源四倍体（3000 万年前的基因组复制事件导致很多基因复制，尽管一些最初复制的基因从那时起就丢失了），其主要的缺点是不容易进行遗传分析。



▶ **热带爪蟾**（*Xenopus tropicalis*）（右上图）

较小的热带爪蟾世代时间相对较短（小于 5 个月），每次产生 1000～3000 个卵，但比光滑爪蟾的卵要小。它在进化上与光滑爪蛙很接近，但它具有二倍体基因组，更易于进行遗传分析。

小鼠



这种模式生物与人类基因组计划关系最密切了（Meisler，1996）。它是具有高度发展的遗传学的哺乳动物，是广泛使用的哺乳动物发育模型。它体型较小、传代时间短，可以进行大规模诱发突变程序和广泛的遗传杂交，它的很多特点有助于基因和表型的定位（框 14.2）。由于人和小鼠编码序列之间相对较高的序列保守程度，所以几乎全部人类基因都可以很容易地找到在小鼠中的同源性基因。染色体大片段在小鼠和人之间保守（图 14.7，图 14.8），因此如果把小鼠基因组的一个区域进行高分辨绘图，那么就可以用这个信息来推测人类基因组的相应部分（反之亦然）。这与医学研究尤其相关，因为种间同源的小鼠和人类突变体通常显示出相似的表型，所以在一个物种中某个疾病基因的定位克隆与其他物种明显相关（节 14.3.5）。对生殖细胞进行预定的遗传修饰（通过胚胎干细胞的转基因和基因打靶技术）来构建小鼠模型，已成为研究基因表达和功能，以及建立人类疾病的小鼠模型的一个强有力的工具。详见 20 章。

框 8.8　用于了解发育、疾病和基因功能的多细胞动物模型（续）

大鼠



大鼠，比小鼠大得多，多年来一直是生理学、神经学、药理学和生物化学分析的首选哺乳动物。它们也可以提供人类复杂的血管和神经疾病的遗传模型系统，如高血压和癫痫（由于各种原因，没有这些疾病的小鼠模型）。而实验室中对大鼠的遗传学研究落后于小鼠很多，部分是因为喂养它们的成本相对较高，而且通过基因打靶修饰大鼠生殖细胞目前还很困难。但最近人们构建了其高分辨率的遗传和物理图，并获得了其基因组序列。

8.4.4　后生生物基因组计划主要集中于发育和疾病模型

秀丽新小杆线虫测序计划的成功开创了一个令人信心百倍的新时代。为了增加少数几个经典的动物基因组计划（例如黑腹果蝇），人们根据不同目的又开展了许多新计划：

- 基础研究——例如为了完全了解有价值的发育模型；
- 商业方面——例如农场动物的基因组计划；
- 医学方面——例如为了了解疾病模型，以及寄生线虫和蚊子（疟疾载体）等疾病载体的致病机制。

黑腹果蝇基因组计划

该计划最初大部分是由加利福尼亚大学 Berkeley 实验室和欧洲实验室协作组共同合作进行的，但后来主要是由私人公司 Celera 参与进行。Adams 等（2000）公布的序列并不是全部 165Mb 基因组序列，但是几乎包含了所有近 120Mb 的常染色质部分，该常染色质部分包含了绝大多数基因。2000 年 8/9 月份发布的 3.0 版本除去了常染色质序列中的大多数间隙（http：//www.fruitfly.org/sequence/index.html）。

最初报道的果蝇基因组有 13601 个编码多肽的基因，基因密度约为每 9kb 一个。当参考最初报道的秀丽新小杆线虫——一种细胞数目大约只有黑腹果蝇 10％的简单生物——有 19000 个编码多肽的基因时，这么少的基因数目是令人吃惊的。为了找出黑腹果蝇和其他远支果蝇种系的关系，最近又启动了其他基因组计划，尤其是拟盲果蝇（*D. pseudoobscura*）计划。

冈比亚按蚊基因组计划

冈比亚按蚊是一种传播致死性疟疾寄生虫——镰型疟原虫的蚊子。主要在 U. S. NIH-NIAID 和法国科研部的赞助下，Celera 基因组公司、法国国家测序中心 Genoscope 和基因组研究所（TIGR）以及几个大学的研究团体，共同对冈比亚按蚊基因组进行测序（Holt *et al.*,2002）。其序列与镰型疟原虫的序列一起公布（节 8.4.2），为治疗疟疾提供了新的空间。

小鼠（小家鼠，*Mus musculus*）基因组计划

由于小鼠的各种特点，所以它提供了与人类基因组计划最相关的基因组模型（框 8.8），并期望含有基本相同数目的基因。由政府出资的小鼠基因组测序工作开始于 1999 年，最初预计到 2005 年前提交一个序列草图。但是，当私人公司 Celera 随后宣布将在两年内测出小鼠基因组序列时，人们做出了一致性尝试来加快公共测序进程。为了实现此目标美国国立卫生研究院和英联邦 Wellcom Trust 以及三家私人公司（GlaxoSmithKline，Merck 基因组研究院和 Affymetrix 公司）共同组成了**小鼠基因组测序协作组**（Mouse Genome Sequencing Consortium，MGSC）。MGSC 测序工作分派给三个主要的参与人类基因组计划的中心：Wellcom Trust Sanger 研究所，Whitehead 研究所/MIT和 Baylor 医学院。

2001 年 4 月，Celera 公司宣布他们竞争的小鼠测序计划已经完成了六倍覆盖的小鼠基因组，包括来自三个小鼠种系的序列：129X1/SvJ，DBA/2J 和 A/J。Celera 的测序数据不是免费获得的，反而通路仅限于那些准备花费高额预定费用的客户。2002 年 5 月，MSGC 公布他们已经完成了七倍覆盖的 C57BL/6J 小鼠基因组，即完成了 96%，并且立即把结果上传到互联网，早于 2002 年 12 月按惯例发表的文章（小鼠基因组测序协作组，2002）。

Ensembl 计划（Ensembl Project）由欧洲生物信息研究所、Wellcome Trust Sanger 研究所（http：//www. ensembl. org/Mus_musculus）及加利福尼亚大学 Santa Cruz 分校（http：//genome. ucsc. edu/cgi-bin/hgGateway？ db＝mm2）共同完成，由于这个计划，可以得到小鼠基因组序列数据的注解版本。来自最初序列草图的数据显示，小鼠基因数量与人类基因数量很近似（约 30000 或略少于 30000 个——见小鼠基因组测序协作组，2002），但也有一些令人感兴趣的差别（节 12.4.1）。这些数据不仅为小鼠基因及其基因组结构提供重要数据，也在确定序列变异和比较基因组学研究中很有价值，例如与人类基因组序列的比较，就非常有助于确定高度保守序列（不仅是编码 DNA 也包括调节及其他一些序列）以及了解基因组进化（节 12.3.2）。

其他后生动物基因组计划（表 8.4）

- **鸟类基因组计划**。由于其作为发育模型的重要性，最初的研究对象是**鸡**（框 8.8）。
- **鱼类基因组计划**。到 2002 年末，已经得到了**河豚鱼**（pufferfish）（一种紧凑的脊椎动物基因组的模型）和**斑马鱼**（zebrafish）（一种发育的模型，也是一种日益重要的疾病和基因功能的模型）的大部分基因序列（表 8.4 和框 8.8）。

表 8.4　重要的后生动物（多细胞动物）基因组计划（进一步阅读见电子参考文献）

动物类型	物种(通用名)	基因组大小	合作中心[a]
海鞘	海鞘(*Ciona intestinalis*)(海鞘)	200Mb	美国 DoE 联合基因组研究所
鸟类	原鸡(*Gallus gallus*)(鸡)	1200Mb	华盛顿大学
鱼类	斑马鱼(*Brachydanio rerio*)(斑马鱼)	1900Mb	Wellcome Trust Sanger 研究所
	河豚鱼(*Fugu rubripes*)(河豚鱼)	400Mb	国际 Fugu 基因协作组
	黑青斑河豚(*Tetraodon nigroviridis*)(河豚鱼)	350Mb	巴黎 Genoscope
蛙类	热带爪蟾(*Xenopus tropicalis*)	1700Mb	美国 DoE 联合基因组研究所
昆虫	蜜蜂(*Apis mellifera*)(蜜蜂)	270Mb	Baylor 医学院
	埃及伊蚊(*Aedes aegypti*)(黄热病蚊)	780Mb	TIGR(基因组研究所)
	冈比亚按蚊(*Anopheles gambiae*)(疟疾蚊，malaria moquito)	278Mb	国际性(主要是 Celera 和 Genoscope)
	黑腹果蝇(果蝇;euchirom. Region)	165Mb	Celera/果蝇基因组中心/Howard Hughes 医学院
	拟盲果蝇(*Drosophila pseudoobscura*)(euchrom. Region)	125Mb	Baylor 医学院
哺乳动物	牛(*Bos taurus*)(牛 cow)	3000Mb	Baylor 医学院
	家犬(*Canis familiaris*)(狗)	2800Mb	Whitehead/MIT
	小家鼠(*Mus musculus*)(小鼠)	2500Mb	小鼠基因组测序协作组/Celera
	黑猩猩(*Pan troglodytes*)(常见黑猩猩)	3500Mb	Whitehead/MIT 和华盛顿大学
	褐鼠(*Rattus norvegicus*)(大鼠)	3100Mb	Baylor 医学院
线虫	秀丽新小杆线虫	100Mb	Wellcome Trust Sanger 研究所/华盛顿大学
	Caenorhabditis briggsae	80Mb	华盛顿大学
海胆	紫色球海胆	800Mb	Baylor 医学院(可能)

a 涉及的主要测序中心网址如下：

Baylor 医学院基因组测序中心，http://hgsc.bcm.tmc.edu/

Celera，http://www.celera.com/

DoE 联合基因组学院，http://www.jgi.doe.gov/

Genoscope，http://www.genoscope.cns.fr/

TIGR(基因组研究所)，http://www.tigr.org/

华盛顿大学基因组测序中心，http://genome.wustl.edu/

Wellcome Trust Sanger 研究所，http://www.sanger.ac.uk/

Whitehead 学院/MIT 基因组研究中心，http://www-genome.wi.mit.edu/

- **蛙类基因组计划**。如框 8.8 所详述，**爪蟾**（*Xenopus*）是一种很好的发育模型，人们已经约定对**热带爪蟾**（*Xenopus tropicalis*）进行基因组测序，它比目前普遍研究的**光滑爪蟾**（*Xenopus laevis*）更容易进行遗传分析。
- **蝇类基因组计划**。除了果蝇和蚊子计划（节 8.4.4），人们还启动了**蜜蜂**（honey bee）的基因组计划，它令人感兴趣的原因在于：①它有很强的社会习性和独特的行为特点（对神经生物学家很有用）；②与人类健康相关（蜜蜂蜇伤潜在的严重后果，

作为抗生素耐受、免疫、过敏反应等的模型)；③作为传粉者对农业的重要性。

▶ **蠕虫基因组计划**。除了非寄生性线虫计划［已完成的秀丽新小杆线虫计划（节 8.4.3）以及最近其远亲 *C. Briggsae* 的测序计划（节 12.3.2)，Wellcom Trust Sanger 研究所、华盛顿大学和爱丁堡大学又发展了一项主要的合作计划，该计划对大约 20 种**寄生性线虫**（parasitic nematode）的几十万个 EST 进行测序。它们包括**人蛔虫**（*Ascaris lumbricoides*）——人类最常见的寄生线虫，感染约 14.7 亿人（蠕虫迁移通过肺部引发肺炎，堵塞消化道或胆管胰管而致病——见 http：//nematode. net)。

▶ **哺乳动物基因组计划**。最近已经开始如下基因组的测序。

- **黑猩猩**（chimpanzee)，我们的近亲。对猩猩基因组进行测序的兴趣源于它在进化史上与人类非常密切的关系。这些信息可以为我们提供有价值的见识来理解在猩猩中罕见的，或者看起来是因为一些医学条件感染人类而不是猩猩的人类疾病（如 HIV 进展为 AIDS，镰型疟原虫导致疟疾）——见 Cyranoski（2002）和 Olson and Varki (2003)；
- **大鼠**（rat)，一个很有价值的心血管、心理学和生理学研究模型。Baylor 医学院公布了它的基因组序列草图，并于 2002 年 11 月列入 GeneBank。该序列覆盖了基因组的 90%（估计总计为 2800Mb 中的 2560Mb；见 http：//www. hgsc. bcm. tmc. edu/projects/rat/)；
- **牛、绵羊和猪**（cow，sheep and pig）——有价值的农业畜类；
- **犬**（dog)，一个有价值的疾病模型（犬类患有许多种与人类相同的疾病)。犬类也是重要的行为遗传学模型，并广泛用于药理学研究。

▶ **纯海洋动物基因组计划**（simple marine animal genome projects）

- **海胆**（Sea urchin）多年来一直是基础生物学，特别是发育生物学研究重要的模型。它是一种后口动物，因此与脊椎动物和人类相对密切相关（图 12.23)。有很多关于海胆基因表达的信息，因此它是用于了解基因和蛋白如何调节生长和发育的一个有用模型。最近开始了紫色球海胆（*Strongylocentrotus purpuratus*）的基因组计划。
- **海鞘**（ascidians）也称**背囊动物**（tunicates)，就是常说的**海鞘**（sea squirts)，是一种海洋生物，大部分生命都在码头、礁石或船底度过。它属于尾系动物门，但与其他无脊椎动物相比，它们实际上与脊椎动物更密切相关。玻璃海鞘（*Ciona intestinalis*）是任何可用于实验操作的脊索动物（chordate）中基因组最小的，为探查脊索动物系的进化起源提供了一个良好的系统。Dehal 等（2002）公布了玻璃海鞘的基因组序列草图。117Mb 的基因组中含有有约 16000 个基因，与已测序的脊椎动物基因组的比较基因组学分析，提供了关于脊索动物和脊椎动物进化的重要信息。

（尚　超　译）

进一步阅读

Borsani G, Ballabio A, Banfi S (1998) A practical guide to orient yourself in the labyrinth of genome databases. *Hum. Mol. Genet.* **7**, 1641–1648.

Database issue of Nucleic Acid Research (2003) *Nucl. Acid Res.* **31**, 1–516.

Hedges SB (2002) The origin and evolution of model organisms. *Nature Rev. Genet.* **3**, 838–849.

Human Genome *Nature* Issue (15 February 2001). *Nature* **409**, 813–958 (papers are available electronically at the Nature Genome Gateway at http://www.nature.com/genomics/human/).

Human Genome *Science* Issue (16 February 2001). *Science* **291**, 1177–1351 (papers are available electronically at http://www.sciencemag.org/content/vol291/issue5507/index.shtml).

Mouse Genome *Nature* Issue (5 December 2002). *Nature* **420**, 509–590 (papers are available electronically at http://www.nature.com/nature/mousegenome/index.html).

User's Guide to the Human Genome. *Nature Genetics Supplement*, September 2002 (available electronically through the Nature Genome Gateway at http://www.nature.com/genomics/).

Wilkie T (1993) *Perilous Knowledge: the Human Genome Project and its Implications*. Faber and Faber, New York.

Electronic information on the Gene Ontology Consortium at http://www.geneontology.org/

Electronic information on the Human Genome Project (and related projects) can be found at many locations. Useful web sites include the following sites:

- the U.S. National Human Genome Research Institute (NGHRI) at http://www.nhgri.nih.gov/
- *the U.S. National Center for Biotechnology Information (NCBI) at* http://www.ncbi.nlm.nih.gov/
- the Genome Web maintained at the UK Human Genome Mapping Resource Centre at http://www.hgmp.mrc.ac.uk/GenomeWeb/
- The Nature Genome Gateway at http://www.nature.com/genomics/

Electronic information on Genome Projects for micro-organisms and model organisms can be found at a variety of sites, including:

- The TIGR microbial genome database at http://www.tigr.org/tdb/mdb/mdbcomplete.html and http://www.tigr.org/tdb/mdb/mdbinprogress.html
- The European Bioinformatics parasite genome webpage at http://www.ebi.ac.uk/parasites/paratable.html
- The Parasite Genomes Web Site at http://www.rna.ucla.edu/par/
- The U.S. national Human Genome Research Institute's web site at http://www.genome.gov/Pages/Research/Sequencing/Proposals/

Electronic information on Model Organisms includes:

- the NIH model organism site at http://www.nih.gov/science/models
- the Model Organisms Virtual Library at http://www.ceolas.org/VL/mo/
- the NCBI Model Organisms site at http://www.ncbi.nlm.nih.gov/About/model/
- the Model and other organisms glossary at http://www.genomicglossaries.com/content/model_organisms_glossary.asp
- the Tree of Life web project at http://tolweb.org/tree/

Electronic summary lists for all Genome Projects can be found at a variety of sites, including:

- A list of completed and ongoing genome projects compiled by Integrated Genomics at http://ergo.integrat-edgenomics.com/GOLD/
- a list of completed genomes at the European Bioinformatics Institute at http://www2.ebi.ac.uk/genomes/

Human Genome Browsers and Integrated Databases include:

- Ensembl at http://www.ensembl.org
- NCBI Map Viewer at http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map-search
- UCSC Genome Browser at http://genome.ucsc.edu
- The HOWDY integrated database at http://www-alis.tokyo.jst.go.jp/HOWDY

参考文献

Adams MD, Kelley JM, Gocayne JD *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science* **252**, 1651–1656.

Adams MD, Celniker SE, Holt RA *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.

Banfi S, Borsani G, Rossi E *et al.* (1996) Identification and mapping of human cDNAs homologous to Drosophila mutant genes through EST database searching. *Nature Genet.* **13**, 167–174.

Batzoglou S, Pachter L, Mesirov JP, Berger B, Lander ES (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958.

Beck CW, Slack JM (2001) An amphibian with ambition: a new role for Xenopus in the 21st century. *Genome Biol.* **2**, 1029.1–1029.5.

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphism. *Am. J. Hum. Genet.* **32**, 314–331.

Botstein D, Chervitz SA, Cherry JM (1997) Yeast as a model organism. *Science* **277**, 1259–1260.

Brown WRA, Hubbard SJ, Tickle C, Wilson SA (2003). The chicken as a model for large scale analysis of vertebrate gene function. *Nature Rev. Genet.* **4**, 87–98.

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94.

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2017.

Camargo AA, de Souza SJ, Brentani RR, Simpson AJG (2002) Human gene discovery through experimental definition of transcribed regions of the human genome. *Curr. Opin. Chem. Biol.* **6**, 13–16.

Cavalli-Sforza LL, Wilson AC, Cantor CR, Cook-Deegan RM, King MC (1991) Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. *Genomics* **11**, 490–491.

Chumakov IM, Rigault P, Le Gall I *et al.* (1995) A YAC contig map of the human genome. *Nature*.**377**, 175–297.

Clark MS (1999) Comparative genomics: the key to understanding the Human Genome Project. *Bioessays* **21**, 121–130.

Cohen D, Chumakov I, Weissenbach J (1993) A first generation physical map of the human genome. *Nature* **366**, 698–701.

Collins FS, McKusick VA (2001) Implications of the human genome project for medical science. *J. Am. Med. Assoc.* **285**, 540–544.

Collins F, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* **262**, 43–46.

Cox DR, Burmeister M, Proce ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science*, **250**, 245–250.

Craig JM, Bickmore WA (1994) The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7**, 376–381.

Cuthbert AP, Trott DA, Ekong RM *et al.* (1995) Construction and characterization of a highly stable human:rodent monochromosomal hybrid panel for genetic complementation and genome mapping studies. *Cytogenet. Cell Genet.* **71**, 68–76.

Cyranoski D (2002) Almost human. *Nature* **418**, 910–912.

Danesh J, Newton R, Beral V (1997) A human germ project? *Nature* **389**, 21–24.

Davies KE, Young BD, Elles RG, Hill ME, Williamson R (1981) Cloning of a representative genomic library of the human X chromosome after sorting by flow cytometry. *Nature* **293**, 374–376.

Dehal P, Satou Y, Campbell RK (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298**, 2157–2167.

Deloukas P, Schuler GD, Gyapay G *et al.* (1998) A physical map of 30,000 human genes. *Science* **282**, 744–746.

Doolittle RF (2002) Microbial genomes multiply. *Nature* **416**, 697–700.

Dunham I, Shimizu N, Roe BA *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* **402**, 489–495.

Gardner MJ, Shallom SJ, Carlton JM *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511.

Garver KL, Garver B (1994) The Human Genome Project and eugenic concerns. *Am. J. Hum. Genet.* **54**, 148–158.

Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433.

Goffeau A, Barrell BG, Bussey H *et al.* (1996) Life with 6000 genes. *Science* **274**, 546–567.

Greely HT (2001) Human genome diversity: what about the other human genome project? *Nature Rev. Genet.* **2**, 222–227.

Hedges SB (2002) The origin and evolution of model organisms. *Nature Rev. Genet.* **3**, 838–849.

Holt RA, Subramaniam GM, Halpern A *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129–149.

Hudson TJ, Stein LD, Gerety SS *et al.* (1995) An STS-based map of the human genome. *Science* **270**, 1945–1954.

Ichikawa H, Hosoda F, Arai Y, Shimizu K, Ohira M, Ohki M (1993) A *Not*I restriction map of the entire long arm of human chromosome 21. *Nature Genet.* **4**, 361–365.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

International SNP map working group (2001) *Nature* **409**, 928–933.

Knoppers BM (1999) Status, sale and patenting of human genetic material: an international survey. *Nature Genet.* **22**, 23–25.

Kong A, Gudbjartsson DF, Sainz J *et al.* (2002) A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247.

Ludecke HJ, Senger G, Claussen U, Horsthemke B (1989) Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification. *Nature* **338**, 348–350.

McKusick V (1989) HUGO News. The Human Genome Organization: History, Purposes, and Membership. *Genomics* **5**, 385–387.

Meisler MH (1996) The role of the laboratory mouse in the human genome project. *Am. J. Hum. Genet.* **59**, 764–771.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Murray JC, Buetow K, Weber JL *et al.* (1994) A comprehensive human linkage map with centimorgan density. *Science* **265**, 2049–2054.

Olsen GJ, Woese CR (1997) Archaeal genomics: an overview. *Cell* **89**, 991–994.

Olson MV, Varki A (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4**, 20–28.

Pennisi E (1997) Laboratory workhorse decoded. *Science* **277**, 1432–1434.

Perrimon N (1998) New advances in *Drosophila* provide opportunities to study gene functions. *Proc. Natl Acad. Sci. USA* **95**, 9716–9717.

Reboul J, Vaglio P, Tzellas N *et al.* (2001) Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in *C. elegans*. *Nature Genet.* **27**, 332–336.

Schuler GD, Boguski MS, Stewart EA *et al.* (1996) A gene map of the human genome. *Science* **274**, 540–546.

Spradling AC, Stern DM, Kiss I, Roote J, Laverty J, Rubin GM (1995) Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl Acad. Sci. USA* **92,** 10824–10830.

Subramanian G, Adams MD, Venter JC, Broder S (2001) Implications of the human genome for understanding human biology and medicine. *J. Am. Med. Assoc.* **286**, 2296–2307.

Tang CM, Hood DW, Moxon ER (1997) *Haemophilus* influence: the impact of whole genome sequencing on microbiology. *Trends Genet.* **13**, 399–404.

Thomas SM, Davies ARW, Birtwistle NJ, Crowther SM, Burke JF (1996) Ownership of the human genome. *Nature* **380**, 387–388.

Turkewitz AP, Orias E, Kapler G (2002) Functional genomics: the coming of age for *Tetrahymena thermophila*. *Trends Genet.* **18**, 35–40.

van Ommen GJ (2002) The Human Genome Project and the future of diagnostics, treatment and prevention. *J. Inherit. Metab. Dis.* **25,** 183–188.

Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science.* **291**, 1304–1351.

Walter MA, Spillett DJ, Thomas P, Weissenbach J, Goodfellow PN (1994) A method for constructing radiation hybrid maps of whole genomes. *Nat. Genet.* **7**, 22–28.

Waterston R, Lander ES, Sulston J (2002) On the sequencing of the human genome. *Proc. Natl Acad. Sci. USA* **99**, 3712–3716.

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M (1992) A second generation linkage map of the human genome. *Nature* **359**, 794–801.

Wittbrodt J, Shima A, Schartl M (2002) Medaka–a model organism from the far East. *Nature Rev. Genet.* **3**, 53–64

Wood V, Gwilliam R, Rajandream MA *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880.

Zhang MQ (2002) computational prediction of eurkaryotic protein-coding genes. *Nature Rev. Genet.* **3**, 698–709.

第9章　人类基因组的组成

本章内容

9.1　人类基因组的一般组成

9.1.1　人类基因组的概貌

人类基因组（human genome）是用来描述人类细胞全部遗传信息（DNA 含量）的术语。它实际上由两个基因组构成：一个复杂的**核基因组**（nuclear genome），有约30 000个基因，另一个是很简单的**线粒体基因组**（mitochondrial genome），有 37 个基因（图 9.1）。核基因组提供大量的基本遗传信息，其大多数在细胞质的核糖体上特化多肽合成。

线粒体有它们自己的核糖体和极少的编码多肽的基因，在线粒体基因组产生mRNA，并在线粒体的核糖体上被翻译。然而，线粒体基因组仅特化很小部分特异的线粒体的功能；而大部分线粒体的多肽是由核基因编码并在输入线粒体之前在细胞质的核糖体上合成。

人类-小鼠比较研究表明少于 5%的基因组是极其保守的，包括 1.5%专用于编码DNA 和百分比稍高的在非翻译序列内的保守序列、调节元件等（小鼠基因组测序协作组，2002；Dermitzakis *et al*.,2002）。大多数编码 DNA 用来形成 mRNA 和此后的多肽，但有极少部分（至少 5%和可能近 10%）人类的基因特化为非编码（=非翻译的）RNA 基因（RNA gene）。最近已鉴定了多种新 RNA 基因，使我们不得不重新评价RNA 功能。

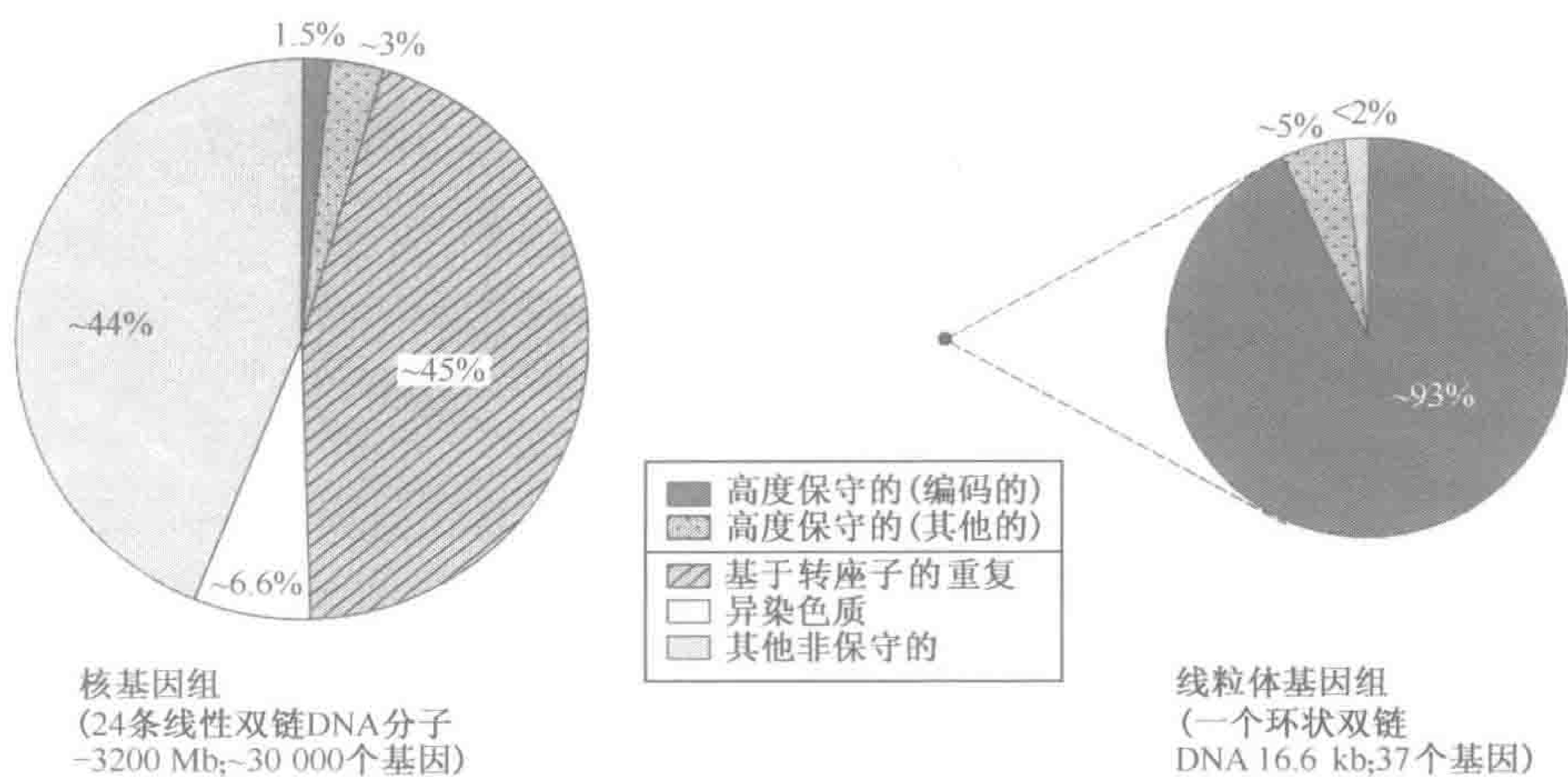


图 9.1 人类基因组的组成

图中的圆点是运用核基因组相同的尺度表示线粒体基因组的相对大小。**注**：在两个基因组的高度保守的 DNA（编码序列，调节序列等）和小部分的高度重复的非编码 DNA 之间都有相当的差别。

编码序列通常属于相关的序列家族（**DNA 序列家族**，DNA sequence family），它们可组成簇，位于一条或多条染色体上或分散其上。这样的复制序列是由不同的**基因复制**（gene duplication）机制产生的，这些机制已发生于进化过程中。基因组的测序首次提供对全基因组复制的评定，并揭示了大量有意义的灵长类特殊的**片段性重复**（segmental duplication）[由于最近期复制的结果，在不同的染色体或单个染色体的不同的区域上发现非常紧密相关的区段序列（节 12.2.5 和图 12.13）；Boiley *et al*.,2002]。

引起基因复制的机制也产生无功能性基因的相关序列，包括**假基因**（pesudogene）和**基因片段**（gene fragment）（节 9.3.6）。还有散在全基因组的 RNA 基因的许多不完整的拷贝，同时某些编码多肽的基因中，也发现了许多有关的假基因：对完成的 21 号和 22 号染色体序列的分析，预测在基因组中总计大约有 20 000 个假基因（Harrison *et al*.,2000；Collins *et al*.,2003）。

正如在其他复杂基因组中那样，人类基因组中很大的成分是由**非编码** DNA（noncoding DNA）构成的。一个相当大的成分是由头→尾（→→→→）的串联重复构成，但其大部分组成为散在的重复序列，这些序列来源于由**反转座子**（retrotransposition）的 RNA 转录物（细胞反转录酶能拷贝 RNA 转录物形成天然的 cDNA，后者能整合到基因组的另一处）。

9.1.2 线粒体基因组由密集包装着遗传信息的一小环状 DNA 双链所组成

线粒体基因组的一般结构和遗传

人类线粒体基因组是由单一类型的环状双链 DNA 所定义的，其完整的核酸序列已经确定（Anderson *et al*.,1981；见线粒体基因图数据库 http://www.mitomap.org），其长16 569 bp，（G+C）占 44%。两条 DNA 链有着十分不同的碱基成分：**重（H）链** [heavy(H) strand] 富含鸟嘌呤，**轻（L）链** [light(L) strand] 富含胞嘧啶。虽然线

粒体 DNA 主要是双链，但有一小段是**三链 DNA**(triple-DNA strand) 结构，由于重复合成重链 DNA 的一个短的片段，7S DNA [见图 9.2 和 Clayton(1992) 关于动物线粒体 DNA 转录和复制的一般综述]。典型的人类细胞含有几千个拷贝的双链线粒体 DNA 分子，而在不同类型的细胞中，其数目能有相当的不同（框 9.1)。

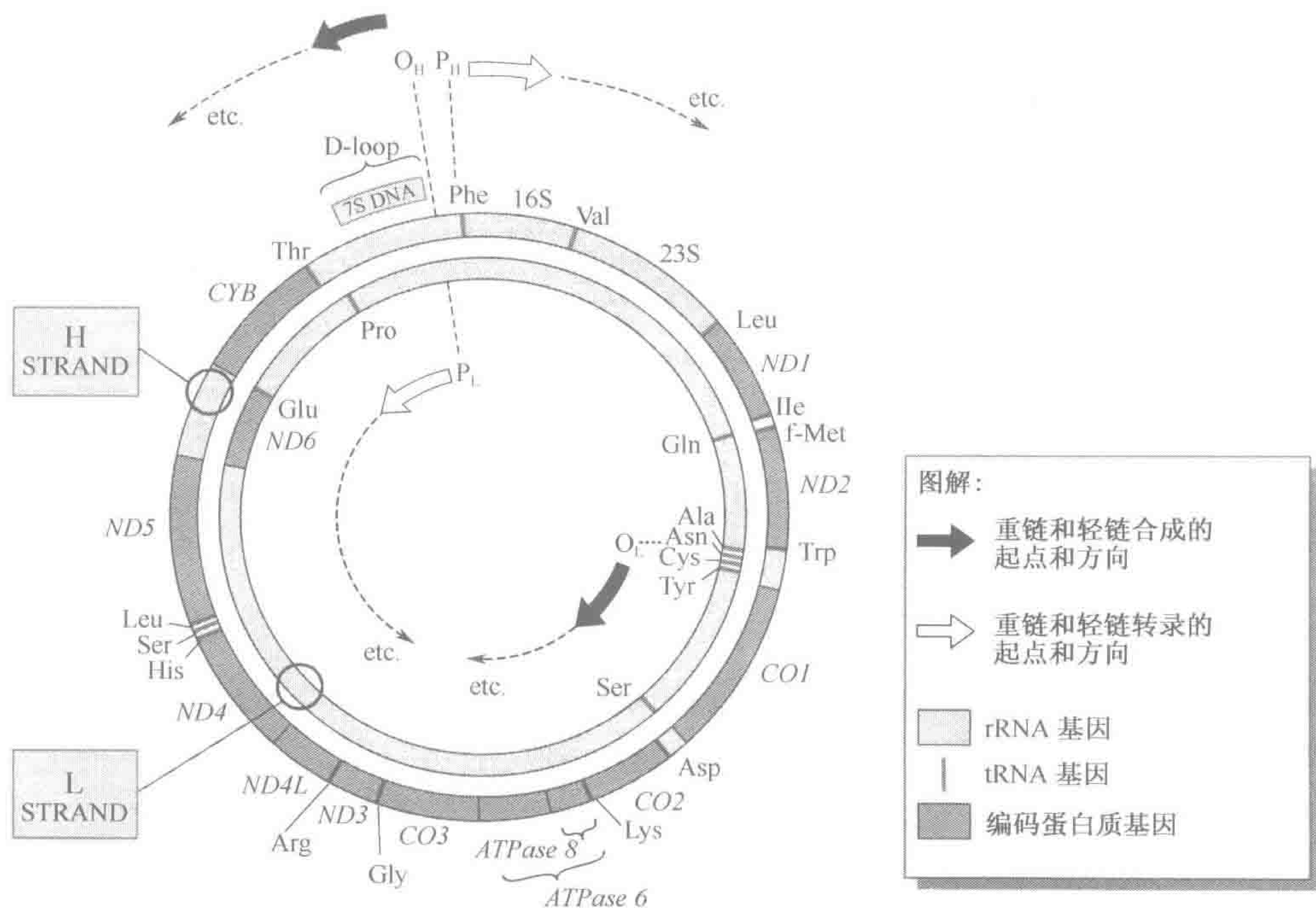


图 9.2 人类线粒体基因组

D loop 是由于一段重（H）链的复制合成的三链结构。重链转录起自 D loop 区域的两个紧密相连的启动子（为明了起见以 P_H表示此组)。从 P_H启动子按顺时针方向一圈转录，而从轻链启动子 P_L则逆时针方向转录。这两种情况下，大的最初转录物切割而产生单个基因的 RNA。所有基因缺少内含子，并以一列重叠基因紧密成簇。ATPase 8 基因部分重叠 ATPase 6 基因（图 9.3)。其他多肽编码基因特化为 7 个 NADH 脱氢酶亚单位（*ND4L* 和 *ND1*～*ND6*)；3 个细胞色素 *c* 氧化酶亚单位（*CO1-CO3*）和细胞色素 *b*（*CYB*)。

框 9.1 人类细胞基因组拷贝数的变化

教科书常称一种有机体细胞的 DNA 含量几乎没有变化，当参照 RNA 或蛋白质含量时，无疑这是真的。不过，在不同类型细胞的 mtDNA 和核 DNA 含量还是有些差别的。

- **线粒体基因组拷贝数的变化**。某些细胞（例如，终末分化的皮肤细胞）缺乏任何线粒体故而没有 mtDNA。mtDNA 拷贝数在其他体细胞中不同，但典型的是在 1000～10 000 之间（例如，淋巴细胞约有 1000 个 mtDNA 分子)。配子是例外的：精细胞有几百个拷贝的 mtDNA 而卵细胞约有 100 000 拷贝，占卵细胞 DNA 的 30%以上。
- **核基因组拷贝数的变化**：有核的（二倍体）细胞表明 DNA 含量没有变化，但不同的**倍体**(ploidy) 意味着某些细胞在 DNA 含量上有本质的不同：无倍体——细胞完全没有 DNA，如许多类型的终末分化细胞，例如红细胞（没有细胞核）和终末分化的皮肤细胞（没有细胞器)；单倍体——卵子和精子有双倍体细胞 DNA 含量的一半；多倍体——某些细胞由于**有丝分裂内复制**

框 9.1　人类细胞基因组拷贝数的变化（续）

(endomitiotic replication）所致自然含有正常染色体组的许多拷贝（那里的细胞进行若干次的DNA复制，但无任何细胞的分裂，例如，肝和其他组织的再生细胞是天然的四倍体，骨髓的巨大百万核细胞能含有多达16倍双倍体细胞的DNA量），或由于**多核体细胞融合**（syncytial cell fusion）所致（例如，由有多细胞核的单细胞产生的多细胞融合的肌纤维细胞—见图3.3）。

在合子形成过程中，精细胞把其核基因组——而不是线粒体基因组——给予卵细胞。因而，合子的线粒体基因组通常只由未受精的卵子最初见到的线粒体所决定。所以，线粒体基因组是母系遗传：男性和女性都从其母亲遗传各自的线粒体而男性不传递线粒体给其后代。因此，线粒体的编码基因或DNA变异体如图4.4所示的系谱模式。在细胞有丝分裂过程中，分裂细胞的线粒体DNA分子是以完全随意的方式分离到两个子细胞中的。

线粒体基因

人类线粒体基因组含有37个基因。其28个基因的重链是有义链；另九个基因的轻链是有义链（图9.2）。在37个基因中，总共24个特化为成熟的RNA产物；22个线粒体tRNA分子和2个线粒体rRNA分子，一个23S rRNA（线粒体的核糖体的大亚基成分）和一个16S rRNA（线粒体的核糖体的小亚基成分）。其余13个是在线粒体的核糖体上合成多肽的编码基因。

由线粒体基因组编码的13个多肽的每一个都是线粒体**呼吸复合物**（respiratory complex）的一个亚单位，此**氧化磷酸化**（oxidative phosphorylation）多链酶参与生产ATP。然而，在线粒体氧化磷酸化系统中，总共约有100个不同的多肽亚单位，而且绝大多数是由核基因编码的（框9.2）。所有其他的线粒体蛋白质都是由核基因编码的，且是在输入线粒体之前，在细胞质的核糖体上翻译的（框9.2，图1.11）。

框 9.2　线粒体基因组的有限自主性

线粒体的成分	由线粒体基因组编码的成分	由核基因组编码的成分
氧化磷酸化系统的成分	13亚单位	＞80亚单位
Ⅰ NADH脱氢酶	7亚单位	＞41亚单位
Ⅱ 琥珀酸辅酶Q还原酶	0亚单位	4亚单位
Ⅲ 细胞色素b-c1复合体	1亚单位	10亚单位
Ⅳ 细胞色素c氧化酶复合体	3亚单位	10亚单位
Ⅴ ATP合成酶复合体	2亚单位	14亚单位
蛋白质合成器的成分	24	～80
rRNA成分	2rRNA	无
tRNA成分	22tRNA	无
核糖体蛋白质	无	～80
其他线粒体蛋白质	无	全部（例如：线粒体DNA和RNA聚合酶加上许多其他的酶，结构和运输蛋白等）

线粒体的遗传密码

线粒体的遗传密码用于解码重链和轻链的转录物以产生总数仅 13 个多肽。这个很小的功能负载已使线粒体的遗传密码漂移于‘通用的’遗传密码（用以维持核基因，由于需要保存 30 000 左右基因的功能）。线粒体有 60 个有义密码子，比核遗传密码少一个，以及 4 个终止密码子。4 个之中的 2 个，UAA 和 UAG，在核遗传密码中也作为终止密码子，而另 2 个是 AGA 和 AGG，二者在核遗传密码中特化为精氨酸（图 1.22）。UGA 编码色氨酸而不是作为终止密码子，而 AUA 特化为蛋氨酸（甲硫氨酸）而不是异亮氨酸。

线粒体基因组编码合成蛋白质需要的全部 rRNA 和 tRNA 分子，而依靠核编码基因以提供所有其他的成分（诸如线粒体核糖体的蛋白质成分，氨基乙酰 tRNA 合成酶等）。因为人类线粒体只有 22 个不同类型 tRNA，每个 tRNA 分子必须能翻译若干不同的密码子。这可能是由于密码子翻译时**第三个碱基摆动**（third base wobble）。在 22 个 tRNA 分子中 8 个有反密码子，它们每个仅能识别第三个碱基不同的 4 个密码子家族，而 14 个可识别一对密码子，它们在头二个碱基位置是相同的，同时第三个碱基则共有一个嘌呤或一个嘧啶。因而，在它们之间，这 22 个线粒体 tRNA 分子能识别总数 60 个密码子〔(8×4)+(14×2)〕。

它们的差异除遗传能力和不同的遗传密码外，线粒体的和核的基因组在其组成和表达的许多其他方面都有不同（表 9.1）。

表 9.1　人类细胞核的和线粒体的基因组

	核基因组	线粒体基因组
大小	3200Mb	16.6kb
不同 DNA 分子数	23(在 XX 细胞中)或 24(在 XY 细胞中);全为线性	一环状 DNA 分子
每个细胞 DNA 分子总数	在双倍体细胞为 46,但依据倍性而异	一般为几千(但有不同,框 9.1)
相关蛋白	若干类组蛋白和非组蛋白的蛋白质	大多没有蛋白质
基因数	约 30 000～35 000	37
基因密度	～1/100kb	1/0.45kb
重复 DNA	超过基因组的 50%(图 9.1)	极少
转录	大量基因是单个转录(单一顺反子转录单位)	从 H 和 L 两个链的多基因共转录(多顺反子转录单位)
内含子	见于大多数基因	缺少
%编码 DNA	～1.5%	～93%
密码子用法	图 1.22	图 1.22
重组	减数分裂时每对同源染色体至少有一次	无重组证据
遗传	在 X 和常染色体上的序列按孟德尔式遗传;在 Y 染色体上的序列为父性遗传	唯一的母性遗传

编码和非编码 DNA

与核基因组不同，人类线粒体基因组是极为紧密的：大约 93%的 DNA 序列是编码

序列。所有 37 个线粒体基因都缺乏内含子并被紧密地包装（平均每 0.45kb 一个基因）。有些基因的编码序列（注意那些编码线粒体 ATP 酶的第 6 和第 8 亚单位的序列）表现某种重叠（图 9.2 和图 9.3），而且在多数情况下，相邻基因的编码序列是由一个或两个非编码碱基连接或分开。某些基因甚至缺乏终止密码子；为了克服这个缺陷，必须在转录后水平引进 UAA 密码子（Anderson *et al*.,1981；见图 9.3 说明）。

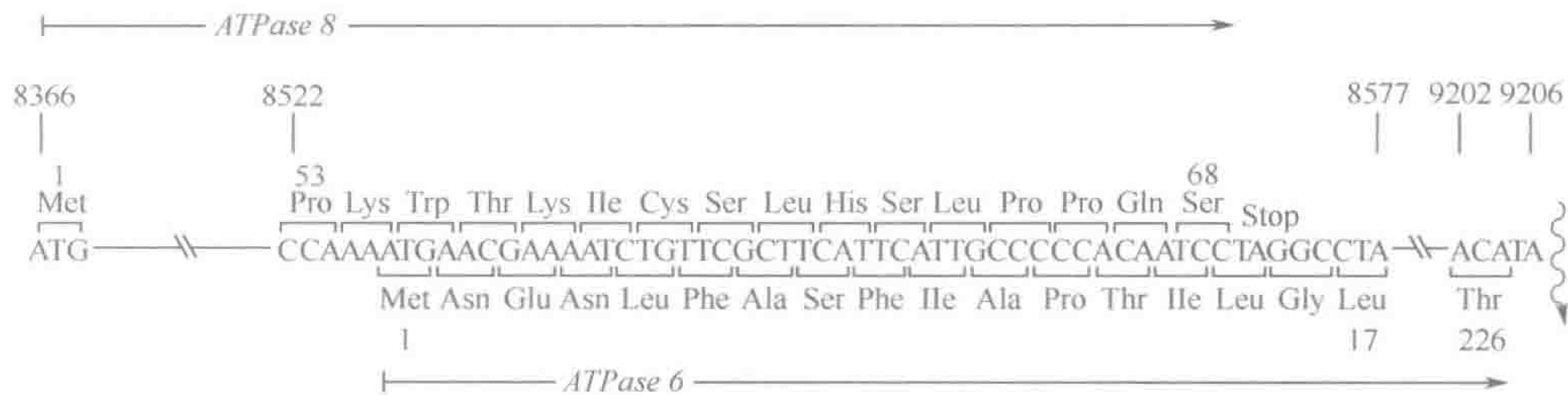


图 9.3　线粒体 ATP 酶亚单位 6 和 8 基因部分重叠并以不同的读框翻译

注：重叠的基因共享一有义链，H 链。共同的编码序列如下：ATP 酶亚单位 8，8366～8569；ATP 酶亚单位 6，8527～9204。ATP 酶亚单位 6 基因的 C 端由转录后引入一 UAA 密码子限定：其转录后的 RNA 在 9206 位后被切割并多聚腺苷化，结果 UAA 密码的头两个核苷酸最终来自 9205～9206 位的 TA，而第三个核苷酸是 poly(A) 尾的第一个 A。其他已知人类基因也有重叠的但通常是由互补链转录的。

唯一缺乏任何已知的编码 DNA 的重要区域是**置换环区域**［displacement(D) loop region］。在此区域内为三链 DNA 结构，是由 H 链 DNA 一短片段复制合成产生的，已知为 7S DNA（图 9.2）。H 和 L 两链的复制是单方向的，并开始于特殊的起点。前者复制的起点是在 D loop，而且仅在大约 2/3 的 H 子链合成之后（通过运用 L 链为模板和置换旧的 H 链），L 链复制的起点才被暴露。此后 L 链的复制以 H 链为模板，按相反方向进行（图 9.2）。D loop 还含有 H 和 L 链转录的主要启动子。不像核基因组的转录那样，每个核基因几乎总是运用各自的启动子分别转录，而线粒体 DNA 的转录都是从 D loop 区的启动子开始，同时，两个不同链以相反的方向连续进行一圈转录以产生大的多基因转录物（图 9.2）。成熟的 RNA 则通过多基因转录物的切割相继产生。

9.1.3　核基因组由与人类 24 条不同染色体相应的 24 个不同的 DNA 分子所组成

人类染色体的大小和结构

人类细胞的细胞核典型地含有 99%以上的细胞 DNA（除某些特化的细胞外，值得注意的是卵细胞，框 9.1）。核基因组分布于 24 个不同类型的线性双链 DNA 分子上，其每一个还有组蛋白和其他非组蛋白与之相结合，构成一个染色体。这 24 个不同的染色体（22 个类型的常染色体和两个性染色体，X 和 Y），通过染色体显带技术易于区分（图 2.15），并主要依其大小，长度和着丝粒的位置分成几大组（表 2.3）。

人类基因组计划测序所选择的 DNA 不是全部的核基因组，而是**常染色质**（euchromatic）部分，近 3000Mb。还有 200kb 以上的**组成型异染色质**（constitutive heterochromatin），为永久浓缩和转录失活区，使总体基因组大小为 3200Mb。因此人类染色体的平均大小约为 140Mb，但在染色体之间有相当的不同，且有不同量的组成型异染

色质（表 9.2）。后者在每个着丝粒约有 3 Mb 片段加上某些染色体上大的成分，包括：端着丝粒染色体 13，14，15，21 和 22 的短臂；Y 染色体的长臂；1 号，9 号和 16 号染色体长臂上的大区域（相当于染色体次缢痕-见表 9.2 和图 2.15）。

表 9.2　DNA 人类染色体的 DNA 含量

染色体	DNA 总量/Mb	异染色质含量/Mb	染色体	DNA 总量/Mb	异染色质含量/Mb
1	279	30	13	118	16
2	251	3	14	107	16
3	221	3	15	100	17
4	197	3	16	104	15
5	198	3	17	88	3
6	176	3	18	86	3
7	163	3	19	72	3
8	148	3	20	66	3
9	140	22	21	45	11
10	143	3	22	48	13
11	148	3	X	163	3
12	142	3	Y	51	27

来自国际基因组测序协作组的摘要数据(2001)，运用这些数字，人类基因组的大小是 3289Mb，但此数据(和个体染色体的总量)已知包括某些人为的重复，而更实际的值可能是约 3200Mb。

人类核基因组的碱基成分

人类基因组序列草图（国际人类基因组测序协作组，2001；Venter *et al*.,2001）表明全基因组常染色质成分的 GC 含量平均为 41%。然而，碱基量在染色体之间有相当的不同，从 4 号和 13 号染色体的 38% GC 含量到 19 号染色体的 49%。同时在染色体上，它也有相当的不同。例如，染色体 17q 的平均 GC 含量在其远端 10.3 Mb 为 50%，而其相邻的 3.9Mb 则降低为 38%。在少于 300kb 区的 GC 含量摆动范围更大，例如从 33.1%到 59.3%。

在染色体显带中 GC 成分和 Giemsa 染色程度之间具有十分明显的关系，例如，定位于深色的 G 带的 98%大片段插入克隆是在 200kb GC 低含量区（平均含量 37%），相反，定位于浅色的 G 带克隆的 80%是在 GC 高含量区（平均 45%）。然而，这些资料的分析并不支持存在精确的**等容线**（isochore），这在成分上曾确定为同源的大尺度区，并根据不同的 GC 成分大约分为 5 个组（国际人类基因组测序协作组，2001）。

核苷酸的某些组合部分有相当大的不同。例如，像其他脊椎动物核基因组那样，人类核基因组明显缺乏双核苷酸 CpG（即在同一 DNA 链的 5′→3′方向胞嘧啶和鸟嘌呤相邻；'P' 表示磷酸二酯键）。按 41% GC 的整体均数估算，单个碱基频率是 C＝G＝0.205，因此双核苷酸 CpG 的预期频率是 $(0.205)^2=0.042$。然而，所观察到的 CpG 频率大约为此值的 1/5。尽管总体上 CpG 缺乏，但一些小的活性 DNA 区的 CpG 密度和预期的一样，并呈显著的非甲基化（CpG 岛，框 9.9）。

9.1.4　人类基因组约有 30 000～35 000 个分布不均的基因，但精确数目未定

人类基因数目

人类基因组基因的总数现在认为是在 30 000～35 000 之间。除 37 个外，所有基因

都位于核基因组内，据此作出一个粗略的估计——每条染色体平均含有 1400 个基因，绝大多数是编码多肽的基因，但极少数（至少 5%或可能约 10%）特化为非翻译的 RNA 基因（节 9.2）。

国际人类基因组测序协作组（2001）和 Verter 等（2001）分别估计为 30 000～40 000和 26 000～38 000 个基因，但有利的预测接近此范围的低限。这一估计大大低于早先根据不完全数据统计的数目（节 8.3.5），但关于基因的精确数目仍相当不确定。首先在鉴定基因上有一般的困难。当基因组草图在 2001 年公布时，确信能鉴定大约 11 000个左右的基因，其他千万个基因是依据计算机对序列分析而预测的。依据计算机预测编码多肽基因曾经是很有帮助，但并不总是可靠的（在鉴定真正外显子时有假阳性且不准确；Zhang，2002）。依据计算机预测 RNA 基因是特别少（节 8.3.5）。

相当低的人类基因数目令人吃惊。毕竟，早先的研究表明很简单的 1 mm 长的秀丽新小杆线虫（*Caenorhabditis elegans*）（由 959 个体细胞构成，并只有一个人类基因组 1/30 大小的基因组）含有 19 099 个编码多肽基因和 1000 以上的 RNA 基因（线虫测序协作组，1998）。基因组复杂性不可能总是与生物学的复杂性相平行［黑腹果蝇（*Drosphilia melanogaster*）实际上要比简单的秀丽新小杆线虫的基因少］，但测序过的无脊椎动物，例如昆虫、线虫、海鞘趋于 14 000～20 000 个基因，而脊椎动物，人、小鼠、斑马鱼等则趋于数目大约为 30 000～35 000 个基因（表 12.4）。这意外的低基因数目亦已作出合理的解释，即根据转录复杂性的大大增加，人们可以预测基因数目变化，譬如说，从 20 000～30 000(Claverie，2001)，同时还由于另外的复杂性，人们能预期在复杂的基因组中选择性剪接频率提高的原因（Maniatis and Tasic，2002；节 10.3.2）。

人类基因的分布

人类基因不是均匀地分布于染色体上的。组成型异染色质区没有基因，但即使在基因组的常染色质部分，基因的密度在染色体区域间、在整个染色体间可有实质的差异。在纯化基因组的 CpG 岛片段与中期染色体杂交后，首次获得了对全基因组基因分布的概括性认识。据此，得出的结论是基因密度在亚端粒区必定是高的。某些染色体（例如 9 号和 22 号染色体）的基因丰富，而其他染色体（例如，X 和 18 号染色体）的基因贫乏（图 8.4）。当报道大约基因组 90%序列草图完成时（国际人类基因组测序协作组，2001），也就确定了不同 CpG 岛密度和不同基因密度的预测。

在 Giemsa 浅带和深带间的 GC 含量差别也反映了基因密度的不同。因为富含 GC 的染色体（例如 19 号染色体）和区（例如 G 浅带）的基因也较丰富。例如，基因丰富的人白细胞抗原（HLA）复合体（在 4Mb 内有 180 基因）位于 6p21.3 浅带，而全部 2.4Mb DNA 几乎只有一单个巨大的基因，抗肌萎缩蛋白基因，位于一 G 深带内。

框 9.3　DNA 甲基化（DNA methylation）和 CpG 岛

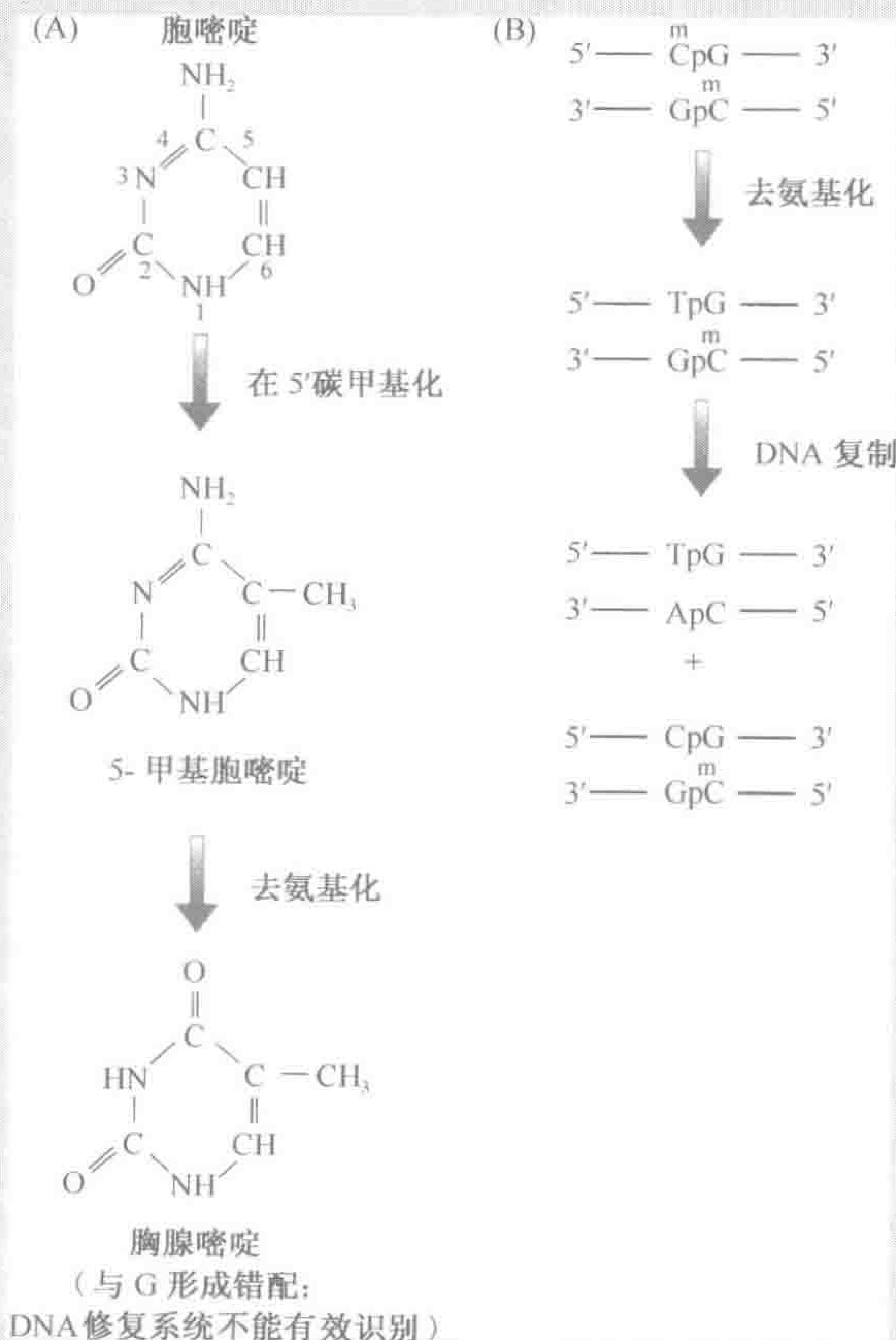
DNA 甲基化可能有不同的生物学作用。对某些物种，如酿酒酵母（*S. cerevisiae*）和秀丽新小杆线虫，似乎完全不发生甲基化；而在许多其他生物，甲基化起着重要作用。细菌 DNA 甲基化大都局限于腺嘌呤和胞嘧啶残基，表现为宿主的防御机制；宿主细胞的限制性内切核酸酶在特殊的识别

框 9.3　DNA 甲基化（DNA methylation）和 CpG 岛（续）

序列识别并切割入侵的（非甲基化的）噬菌体 DNA，而在宿主内的同样序列则特异地甲基化，故而免遭切割（框 5.2）。

在后生动物（多细胞动物）中，DNA 甲基化通常涉及一部分胞嘧啶的甲基化，产生 **5-甲基胞嘧啶**（5-methylcytosine）（C^m）。在黑腹果蝇中 DNA 甲基化总量很低，且多数 5-甲基胞嘧啶见于 CpT 双核苷酸（C^mpT）。在其他动物中，CpG 双核苷酸是胞嘧啶甲基化的常见靶标，通过特异的胞嘧啶甲基转移酶作用形成 C^mpG（图 A）。大多数无脊椎动物的基因组与果蝇不同，具有中度高水平的 C^mpG，它们集中于大的甲基化 DNA 结构域，并被同样大的非甲基化 DNA 结构域分开（**嵌合性甲基化**，mosaic methylation）。

脊椎动物在动物界具有最高水平的 5-甲基胞嘧啶，而这种甲基化分散于全基因组。已知 DNA 甲基化对基因表达有重要的影响，并使特殊基因的表达模式稳定地传递到子细胞（节 10.4.2）。这也说明甲基化为宿主提供了对转座子的一种防御形式（节 10.4.3）。虽然甲基化分散于脊椎动物全基因组，只有百分比很小的胞嘧啶是甲基化（人类 DNA 约为 3%，大多为 C^mpG，但也有百分比很小的 C^mpNpG，这里 N 为任一核苷酸）。



(A) CpG 双核苷酸中的胞嘧啶 5′位碳是甲基化的靶点，形成 5-甲基胞嘧啶
后者自发地去氨基形成胸腺嘧啶（T），它不能由 DNA 修复系统有效识别并趋向存留（然而，非甲基化系统的胞嘧啶去氨基化产生尿嘧啶，可由 DNA 修复系统识别）。(B) 脊椎动物 CpG 双核苷酸逐渐被 TpG 和 CpA 所替代。

5-甲基胞嘧啶在化学上是不稳定的而且易于去氨基化，形成胸腺嘧啶（图版 A）。其他的碱基也易于去氨基化（例如，非甲基化的胞嘧啶易于去氨基化形成尿嘧啶），在进化的长河中，脊椎动

框 9.3　DNA 甲基化（DNA methylation）和 CpG 岛（续）

物 DNA 的 CpG 双核苷酸的数目由于 CpG→TpG(→CpA 于互补链上）的转换慢而稳定，并逐渐减少。虽然脊椎动物基因组 CpG 的总频率低，但有一小段的非甲基化 DNA 以具有正常的和预期的 CpG 频率为特征。这样正常 CpG 密度的岛（**CpG 岛**）较富含 GC（典型的超过 50%GC）并超过几百核苷酸，常成为基因 5′端的标记。在人类基因组序列草图除去高拷贝数的重复非编码 DNA 序列时，能鉴定大约 30 000 CpG 岛（国际人类基因组协作组，2001）。

9.2　人类 RNA 基因的组成、分布和功能

虽然大多数人类基因编码多肽（节 9.3），但有少量重要的基因特化为以非编码的（=非翻译的)RNA 分子为其最终的产物，故称为 RNA 基因（Eddy，2001；Huttenhofer *et al.*,2002；Storz，2002；也见非编码 RNA 数据库 http://biobases.ibch.poznan.pl/ncRNA/）。线粒体基因组 65%（24/37）的基因特化为成熟的 RNA 分子是个例外，但即使在核基因组情况下，也可能有约 3000 个 RNA 基因，为基因总数的近 10%（图 9.4）。

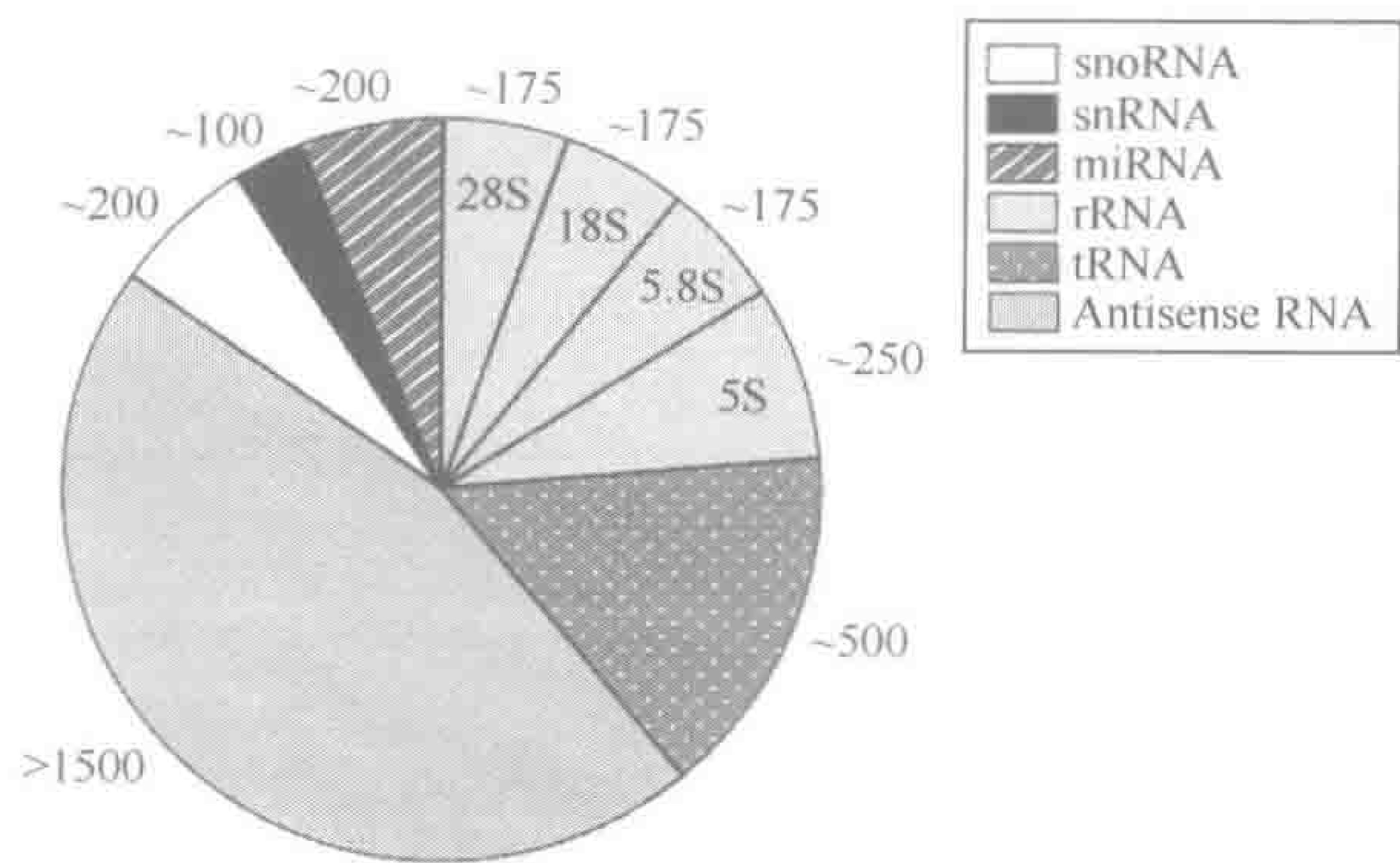


图 9.4　人类 RNA 基因的分类

按 2003 年中期最好的估计，人类 RNA 基因总共超过了 3000 个，分布于图示的不同类型间。

注：（Ⅰ）由于操作原因（见课文），人类基因组序列草图不包括 rRNA 基因簇和由其他资料估计的一定数目；（Ⅱ）由于鉴定 RNA 基因的困难（节 8.3.5），某些类型小 RNA 数，如 miRNA 可能被大大地低估了；（Ⅲ）预测的反义 RNA 基因的数目是出自 Collins 等（2003）的资料，并得到来自小鼠相同分析结果的支持［由 FANTOM 协作组和 RIKEN 基因开发研究组 Phase Ⅰ 和 Ⅱ 小组（2002)］。

现在对 RNA 基因数目的估计可能是保守的（由于在测序的 DNA 中鉴定 RNA 基因的困难性，节 8.3.5）。全面的分析小鼠的转录物（节 9.2.3）和用微阵分析人类 21 号和 22 号染色体的转录物（Kapranov *et al.*,2002）已经说明有比预测的基因数目更多的转录物。此外，特别在由 RNA 聚合酶Ⅲ转录 RNA 基因的情况下，RNA 基因还有许多有关的假基因/基因片段。

与其他细胞基因组相同，多数已知的 RNA 基因致力于形成某些分子，以帮助基因的一般表达过程（图 9.4）。值得注意的是，某些 rRNA 和 tRNA 家族与 mRNA 的翻译有关。许多其他 RNA 家族与 RNA 的成熟有关，涉及其他 RNA 分子的切割和特异碱

基修饰（mRNA、rRNA、tRNA 和其他种类的 RNA）。此外，最近已鉴定了许多属于不同类型的其他 RNA 基因。许多 RNA 基因具有或预期有重要的调节作用，并强调 RNA 分子具有很重要的功能多样性（表 9.3；节 9.2.3）。

表 9.3　人类 RNA 的功能多样性

RNA 分类	举例	功能
(A)**帮助一般基因表达的主要 RNA 类型**		
核糖体 RNA(rRNA)	16S rRNA	线粒体核糖体小亚基的成分(图 9.2)
	23S rRNA	线粒体核糖体大亚基的成分(图 9.2)
	28S、5.8S 和 5S rRNA	细胞质核糖体大亚基的成分(图 10.2)
	18S rRNA	细胞质核糖体小亚基的成分(图 10.2)
转运 RNA(tRNA)	22 种线粒体 tRNA	与线粒体的 mRNA 密码子结合(图 9.2)
	49 种细胞质 tRNA	与细胞质的 mRNA 密码子结合(图 9.4)
小核 RNA(snRNA)	许多，包括：	
(涉及 RNA 剪接)	U1、U2、U4 和 U6 snRNA	大剪接体的成分
	U5 snRNA	大、小剪接体的成分
	U4acat、U6acat、U11 和 U12 snRNA	小剪接体的成分
	U7 snRNA	组蛋白 mRNA 转录终端
小核仁 RNA(snoRNA)	100 以上不同类型：	
(涉及 RNA 修饰和加工)	约 80C/D 盒 snoRNA	rRNA 2′羟基的特殊甲基化位点
	约 15H/ACA snoRNA	由假尿嘧啶形成 rRNA 修饰特殊位点
	U3 和 U8 snoRNA	rRNA 加工
(B)**其他 RNA 类型**(也见非编码 RNA 数据库 http://biobases.ibch.poznan.pl/ncRNA/)		
小 RNA	可能至少 200 种	很小(～22ntds)调节 RNA 分子(节 9.2.3)
与 X 染色体失	*XIST* RNA	见节 10.5.6
活有关的 RNA	*TSIX* RNA	见节 10.5.6
与印迹相关的 RNA	许多，例如：*H19* RNA	见图 10.24 的一些例子
神经系统特异的 RNA	例如：BC200 RNA	?
反义 RNA	可能 1500 种左右的类型	例如：*HOXA11*，*MSX1* 等(图 10.24)
其他	端粒酶 RNA	端粒酶的成分(节 2.2.5)
	PCA3 RNA	前列腺癌抗原 3
	PCGEM1 RNA	在前列腺癌极高表达
	SRA1 RNA	某些类固醇受体的特异共激活子
	TTY2 RNA	睾丸特异家族
	7SK RNA	RNA 聚合酶Ⅱ伸延的负转录调节子
	7SL RNA	转运蛋白质的信号识别颗粒的成分

9.2.1　总共约 1200 个人类基因编码 rRNA 或 tRNA 且多数构成大的基因簇

核糖体 RNA（rRNA）基因

大约 700～800 个人类 rRNA 基因，大多构成串联重复簇和许多有关的假基因。以串联排列出现的同源多基因家族是人类基因组序列草图中未描述过的（由于在构建 BAC 文库时限制酶的选择并用低复杂性指纹表现串联重复 DNA 以延迟 BAC 测序决定

的)。所以，不能从人类基因组序列草图推断 rRNA 基因的精确数目。

除 16S 和 23S 线粒体 rRNA 分子外，还有 4 种类型的细胞质 rRNA，三种与核糖体大亚基有关（28S、5.8S 和 5S rRNA)，一种与核糖体小亚基（18S rRNA）有关。其中 28S、5.8S 和 18S rRNA 由一**单个转录单位**（single transcription unit）（图 10.2）编码，它们构成 5 个簇，每个有 30～40 串联重复，位于人类 13、14、15、21 和 22 号染色体的短臂。

5S rDNA 基因也表现为串联排列，其中最大的在染色体 1q14-42，紧靠端粒。在这些序列中还有 200～300 个真正的 5S 基因，但它们似是许多散在的假基因。对这些细胞质 rRNA 基因重复性的主要理论的说明是依据基因剂量；由于具有大量的这类基因，细胞才能满足细胞质的核糖体进行蛋白合成的巨大需求。

转运 RNA（tRNA）基因

除 22 个线粒体 tRNA 基因外，2001 年发表的人类基因组序列草图揭示了总数有 497 个编码细胞质 tRNA 分子的核基因和 324 个推测衍生的 tRNA 假基因。因此，人类特化的细胞质 tRNA 基因看来要比蠕虫（584）的少，但比果蝇（284）的多。在后生动物中，tRNA 基因数目并不和有机体的复杂性有关，而是与胚胎发育的一定组织或阶段对 tRNA 丰度的特殊需求有关［例如，青蛙（*Xenopus laevis*）有大的卵细胞，其每个一定要负载多至 40ng 的 tRNA。这样高的需求要有上千个 tRNA 基因来满足］。

框 9.4　真核细胞的细胞质 tRNA 的反密码子的特性

正如翻译线粒体密码子的情况（节 9.1.2)，**第三碱基摆动**（third base wobble）意味着在细胞质 mRNA的密码子和识别它们的 tRNA 反密码子间不存在 1∶1 的对应关系。在这种情况下，在第三碱基位有 C 或 U 不同的选择性密码子能被单一反密码子所识别。关于细胞质 mRNA 密码子的解码规则是：

- ‘两密码子框’中的密码子（有 U/C 为末端的密码子在由 A/G 为末端时编码不同的氨基酸)。这里 U/C 摆动位置典型地由在 tRNA 反密码子的 5′碱基位置的一个 G 所解码。因此，对于 Phe（苯丙氨酸)，不存在一个具有**A**AA 反密码子的 tRNA 去匹配 UU **U**密码子，而是**G**AA 反密码子能识别 mRNA 的 UU **U**和 UU **C**密码子（见图 9.5)。
- ‘四密码子框’中的非甘氨酸密码子（在摆动位置 U、C、A 和 G 密码子全都编码相同的氨基酸而不是甘氨酸)。这里的 U/C 摆动位置是由在 5′位反密码子一**次黄嘌呤核苷**（inosine）(1）解码的（次黄嘌呤核苷由腺嘌呤的转录后修饰而产生：腺嘌呤核苷的 6 碳氨基由-C＝O 羰基所替代)。例如，四密码子缬氨酸框的 GU **U**和 GU **C**密码子是由一有**A**AC 反密码子的 tRNA 解码的，这无疑将**A**AC 修饰为**I**AC。次黄嘌呤核苷除与 C 和 U 配对外，还能与 A 配对（因而**I**AC 反密码子能识别 GU **U**、GU **C**、GU **A**中的每一个)。为避免可能的翻译错误，在反密码子 5′碱基位具有次黄嘌呤核苷的 tRNA 不能用于两密码子框；
- 甘氨酸密码子。GGU 和 GGC 密码子是由**G**CC 反密码子而不是预期的**I**CC 反密码子解码的。

只有 16 个反密码子是解码 32 个末端为嘧啶的密码子所需要的。因而，最小一套反密码子是 61（不同的有义密码子数）减去 16 后为 45。然而还有一个特异的 tRNA 携带一针对密码子 UGA 的反密码子（其正常功能是作为一终止密码子)。在高硒条件下，此 tRNA 将解码 UGA，仅在很少情况下将第 21 个氨基酸，**硒代胱氨酸**（selenocysteine)，插入在**“硒代蛋白质”**（selenoprotein）的一个选

框 9.4　真核细胞的细胞质 tRNA 的反密码子的特性（续）

择基中（全部基群都有氧化还原作用；哺乳动物硫氧还蛋白还原酶和谷胱甘肽过氧化酶最常见于硒代蛋白质中）。

497 个细胞质 tRNA 基因按其反密码子的特性分为 49 个家族，虽然通用的遗传密码提供 61 个不同有义密码子，但是需要由 tRNA 分子的反密码子来识别。密码子第三位碱基的摆动意味着当该位的碱基是嘧啶（U 或 C）时，一单个反密码子能与这两个选择密码子配对。为了翻译选择性的人类密码子，反密码子的挑选遵循真核细胞的细胞质 tRNA 的一般规律（框 9.4）。据此，人们预测人类总共有 46 种不同型的 tRNA，而不顾第三个碱基摆动的通则。这样的三对密码子 AU（U/C）、UA（U/C）、AA（U/C）似乎是由两个反密码子的每一个来识别的，因此，还有额外的三类 tRNA（图 9.5）。人类 tRNA 基因与氨基酸频率只有很粗略的对应关系。

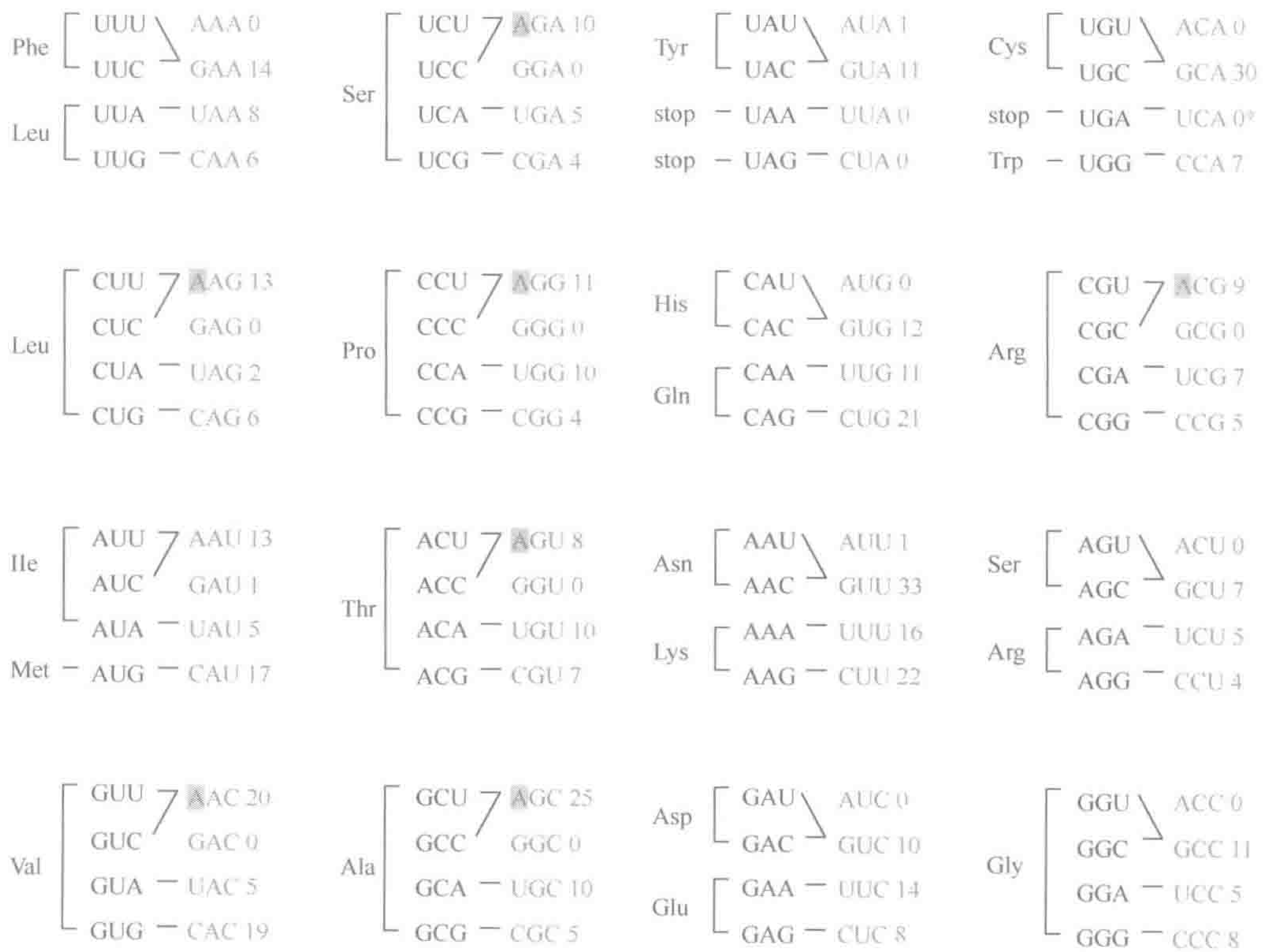


图 9.5　按反密码子分类的人类 tRNA 基因数

密码子以黑线与右侧（未修饰的）反密码子相连。以 V 形相连以一 U 或 C 为末端的选择性密码子，由于第三个碱基摆动，它们能由单个反密码子解码。几乎每个反密码子的数目就是具有其反密码子的人类编码 tRNA 基因的数目。因此，例如，在顶部左侧所见的 UUU 苯丙氨酸密码子不是由 AAA 反密码子解码的，因为没有携带这样一个反密码子的 tRNA 基因。有阴影的腺嘌呤几乎一定被修饰为次黄嘌呤核苷（框 9.4）。**注**：（Ⅰ）尽管预备了第三碱基摆动，单个基因像是编码 tRNA，它具有可能不曾期望是所需要的反密码子（AUA，AUU 和 GAU）；（Ⅱ）星号旁边的反密码子 UCA 表示有一个携带此反密码子例外的 tRNA，这个反密码子偶尔能解译一小组的 UAG 密码子作为硒代胱氨酸替代终止密码子（框 9.4）。修改自 International Human Genome Sequencing Consortium(2001). Nature 409，860～921，经 Nature Publishing Group 允许。

虽然 tRNA 基因像是分散于全基因组（除 22 号和 Y 染色体外，他们可见于其他所有染色体上），而且显著的簇集。其一半以上（280/497）位于 6 号染色体（它有 140 个 tRNA 基因——包括几乎所有不同类型的 tRNA 基因-只在 6p2 的 4Mb 区域内），或位于 1 号染色体［这里有许多天冬酰胺（Asn）和谷氨酸（Glu）tRNA 基因松散地簇集］。此外，许多其他簇集的 tRNA 基因，例如，30 个半胱氨酸（Cys）tRNA 中的 18 个可见于 7 号染色体狭小的 0.5Mb 区内。

9.2.2　小核 RNA 和小核仁 RNA 由大多散在、中等大小的基因家族编码

除 rRNA 和 tRNA 外，还涉及两类帮助一般基因表达的主要 RNA：小核 RNA（snRNA）和小核仁 RNA(snoRNA)。它们是由邻近的 100 个基因（snRNA）或稍多于 100 个基因（snoRNA）的家族所编码。尽管它们是散在的，但也表现为某些簇集的亚家族。

小核 RNA(snRNA) 基因

小核 RNA(small nuclear RNA，snRNA）分子的异源性集合包含许多富含尿嘧啶核苷酸并因而被命名。例如，U3 snRNA 以第三富含尿嘧啶的小核 RNA 而被分类。有些是**剪接体 RNA**（spliceosomal RNA)，为发挥大、小剪接体功能所需要（表 9.3）并由 80 个以上基因的家族所编码。70 多个这样的基因特化为 snRNA 用于大剪接体。它们包括 44 个已鉴定的特异性 U6 snRNA 基因和 16 个特异的 U1 snRNA 基因。

虽然有某些特殊簇集于 U1 和 U2 snRNA 家族的证据，但由于为获得基因组序列草图而选择了 BAC 方式（上述），所以在草图序列中描述的过少。早先已知 U2 RNA 基因位于 *RNU2* 基因座，即在 17q21-q22 上，为 6.1Kb 单位几乎相同的串联排列，其重复单位的数目极为不同（从 6～30 以上重复）。U1 RNA 基因大约 30 拷贝簇集于 1p36.1 的 *RNU1* 基因座，但此簇组成是松散和不规则的。还有大量有关非功能的序列（假基因、基因片段等）；例如 1135 U6 snRNA 相关序列已鉴定与序列草图中序列相一致（国际人类基因组测序协作组，2001）。

小核仁 RNA(snoRNA) 基因

一个**小核仁 RNA**(snoRNA）的大家族大多在核仁中用来直接或指导 rRNA 中特殊位点碱基修饰（Smith 和 Steitz，1997；Filipowicz，2000），但已知它们也对其他稳定的 RNA（包括 U6 snRNA）进行碱基修饰。有两个亚家族，C/D 框 snoRNA 多涉及指导特殊位点 2′-O-核糖体甲基化（在 rRNA 上有 105～107 种这样的甲基化）。H/ACA snoRNA多涉及指导特殊位点假尿苷化（在此尿嘧啶被异构化产生**假嘧啶**（pseudouridine)，为最常见的修饰碱基；rRNA 需要 95 种不同的假尿苷化）。

单一 snoRNA 特化一个或最多两个这样的修饰。snoRNA 基因常见于其他基因的内含子内，同时，虽然大多 snoRNA 基因像是单拷贝和散在的，但是已知某些大基因簇包括两个发现在 15q 上大的 *SNURF-SNRPN* 转录单位。后一基因是父性印记并在脑中表达，而且认为在 Prader-Willi 综合征中起着重要作用（图 10.24 和其中参考文献）。

9.2.3　微 RNA 和其他新的调节 RNA 正挑战关于 RNA 功能范围的偏见

由于蛋白质的广泛功能及其在调节基因表达中的重要性，教科书通常以其作为基因表达的终点而强调蛋白质的重要性。一般认为 RNA 不太重要［由 Venter 等发表的 Celera 基因组序列草图论文（2001）没提供关于人类 RNA 基因的任何分析!］。然而，最近几年许多发现令人不得不从根本上重新评价 RNA 的功能。1982 年，对某些 RNA 分子具有催化功能［起到**核酶**（ribozyme）的作用］的真正认识已导致在若干其他类型 RNA 中发现了催化功能。它们包括 rRNA［最近 X 晶体衍射资料表明肽键的形成是由 rRNA，也由 snRNA(Valadkhan and Manley，2001）催化的而不是蛋白质（Nissen *et al.*,2000)］。

表 9.4　按氨基酸特化的人类细胞质 tRNA 基因家族基因的分布

氨基酸	频率*	相应的 tRNA 基因数	氨基酸	频率*	相应的 tRNA 基因数
丙氨酸	7.06%	40	赖氨酸	5.65%	38
精氨酸	5.69%	30	蛋氨酸	2.23%	17
天冬氨酸	4.78%	10	苯丙氨酸	3.75%	14
天冬酰胺	3.58%	34	脯氨酸	6.10%	25
半胱氨酸	2.25%	30	硒代半胱氨酸	<0.01%	1
谷氨酰胺	4.63%	32	丝氨酸	8.00%	26
谷氨酸	6.93%	22	苏氨酸	5.31%	25
甘氨酸	6.62%	24	色氨酸	1.30%	7
组氨酸	2.56%	12	酪氨酸	2.76%	12
异亮氨酸	4.43%	19	缬氨酸	6.12%	44
亮氨酸	9.95%	35			

* 人类全蛋白质组平均频率。

对许多其他关键的 RNA 分子已有了很好的研究，包括**端粒酶 RNA**（telomerase RNA）（节 2.2.5）和**信号识别颗粒**（signal recognition particle）的 SRP RNA（亦称 7SL RNA）（为了输出，核糖核蛋白复合体识别指定的蛋白质信号序列，给予蛋白质穿过细胞膜的通道）。最近，已鉴定了许多令人兴奋的、新的人类 RNA 分子，具有已知的或预期的调节作用。这是一连续的过程也是至今对哺乳类转录物（从小鼠基因组）最透彻的分析，并表明很大比例的转录物将是非编码 RNA(FANTOM 协作组和 RIKEM 基因组开发研究组 phase Ⅰ & Ⅱ 小组，2002)。

微 RNA：新的小 RNA 调节分子

微 RNA(MicroRNA，miRNA）是很小的（长约 22 核苷酸）RNA 分子，能作为其他基因的反义调节子而发挥作用（Ambros，2001；Gottesman，2002)。它们来自较大的～70 核苷酸长的前体，含有一反向重复序列，使之形成双链**发夹** RNA（hairpin RNA)。这样的发夹前体 RNA 由一类核糖核酸酶Ⅲ切割（对双链 RNA 有特异性)，该酶已知为 dicer[a]。最先在动物中描述了这样的序列，lin-4 和 let-7 RNA，通过对秀丽新

小杆线虫的遗传分析，鉴定为**小瞬时 RNA**(small temporal RNA，stRNA)。lin-4 和 let-7两者受发育的调节，并控制各种发育程序自身。它们作为反义调节子，通过与靶基因的 mRNA 3′UTR 的互补序列相结合，抑制翻译并因而抑制靶基因蛋白质的合成（图 9.6A)。其他 miRNA，例如在植物中也被证明是发育调节子，这一发现加上 miRNA 强有力的进化保守性以致使人们预测哺乳动物的 miRNA 也有相似的功能。

图 9.6　miRNA 是很短（21～22 核苷酸）的可作为反义调节子起作用的 RNA

(A) miRNA **前体的结构**。秀丽新小杆线虫 *lin-4* 和 *let-7* stRNA 属于小 RNA 一类，并表明与哺乳动物小 RNA 相似，诸如人类的 mir-26a 小 RNA。成熟的小 RNA(miRNA）序列（阴影）来自具有反向重复的前体（它由分子间的氢键连接形成一发夹 RNA)，由一类 RNaseⅢ 核酸酶，已知为 dicer 切割发夹 RNA。有时两个不同的 miRNA 来自同一前体。(B) **由 *lin-4* 和 *let-7* RNA 进行反义调节**。来自靶基因的 miRNA 受 *lin-4* 和 *let-7* RNA 调节是在于其 3′UTR 区与这些 miRNA 有十分显著的互补性，碱基配对通常并不完善，正如在 *let-7* RNA 和 *lin*-413′UTR序列的两个靶序列之间预测碱基配对的情况。引自 Banerjee 和 Slack(2002). Bioessays 24，119～129，经 John Wiley & Sons 的子公司 Wiley-Liss 公司同意重印。

曾用不同的方法鉴定哺乳动物的 miRNA(Gottesman，2002 的一个总结)，导致最近在人类和小鼠中鉴定了一些新的小 RNA(Lagos-Quintana *et al*.,2001，2002；Mourelatos *et al*.,2002)。在写作此书时（2003 年中期)，估计已有 200 左右的人类 miRNA。虽然它们分散在许多染色体上，但有证据表明某些是簇集的，值得提到的是一簇至少包含 7 个 miRNA 基因位于 13 号染色体 0.8kb 区内（Mourlatos et al.,2002)。正如所料，哺

乳动物的 miRNA 具有重要的调节功能，现正积极努力地进行靶基因的鉴定。

a注：这类核酶的功能是形成保守的**遗传监视系统**（genetic surveillance system）的一部分，在与特异 mRNA 相当的双链 RNA 起反应时，该系统能降解某一特异 mRNA。它也能用于某些实验分析，正如 RNA **干扰**（RNA interference，RNAi）在细胞内特异地失活靶基因，即通过人工诱导转基因产生一 22 核苷酸长的反义 RNA 分子（即 siRNA＝short interfering RNA）切割长双链 RNA。此 siRNA 分子能与相当于诱导转基因的内源性基因 mRNA 的碱基配对，因此可特异地抑制表达（节 20.2.6）。

编码中等到大尺度的调节 RNA 分子的基因

详细列入已知的或预期的具有调节功能的中等到大尺度的非编码 RNA 基因的数目正在增加［许多但并非全部都是**“非编码 mRNA”**（noncoding mRNA）分子，因为它们在 RNA 聚合酶Ⅱ的作用下经历带帽和多聚腺苷化[b]］。它们包括特化为 *7SK RNA* 的基因，一个 RNA 聚合酶Ⅱ延伸的负转录调节子（Yang *et al*.,2001）；*SRA1*（steroid receptor activator）*RNA*，作为某些类固醇受体的一种特异的共激活因子（Lanz *et al*., 1999）；*XIST*，它对 X-染色体失活非常重要（节 10.5.6）。

某些中等到大尺度的调节**反义** RNA（antisense RNA）已知还包括调节 *XIST* 的 *TSIX* 反义转录物，各种调节印记基因的反义转录物（节 10.5.5；见图 10.24 的一些例子）和许多其他的反义转录物（Lehner *et al*.,2002）。虽然这种反义转录物的总数在人类基因组中仍不确切知晓，但是对人的 22 号染色体上的基因进行了重新评定，已鉴定 16 个可能的反义 RNA 基因，认为在人类基因组中可能约有 1500 个左右的反义 RNA 基因（Collins *et al*.,2003）。为了支持这一评定，FANTOM 协作组和 RIKEN 基因组开发研究组 PhaseⅠ & Ⅱ小组（2002）进行了很全面的研究，使用最保守的估计，已预测了小鼠具有几百个反义 RNA。

b注：经过比较 28S、18S 和 5.8S RNA 由 RNA 聚合酶Ⅰ转录；5S RNA、tRNA、snoRNA 和 miRNA 由 RNA 聚合酶Ⅲ转录；令人惊奇的是 snRNA 为混合转录：有的由聚合酶Ⅲ，而有的由聚合酶Ⅱ转录。

9.3　人类编码多肽基因的组成、分布和功能

9.3.1　人类基因的大小和内部组成存在巨大的变异

大小多样性

在细菌这样简单的有机体中，基因的大小较为相似，且通常很短。在复杂有机体中，特别是人类基因组的基因大小变化相当大（图 9.7）。某些巨大的人类基因意味着其转录要耗费时间，2.4Mb 的肌萎缩蛋白基因大约需要 16 小时（Tennyson *et al*., 1995）。虽然基因与其产物大小间有着直接关系，但也有某些明显的异常。例如，**载脂蛋白**（apolipoprotein）B 的 4563 个氨基酸仅由一 45kb 基因编码，而由巨大的 2.4Mb 肌萎缩蛋白基因编码的最大蛋白质只有 3685 个氨基酸。

外显子-内含子组成的多样性

极少数没有内含子的人类基因（表 9.5）一般是小的。而有内含子的那些基因其大

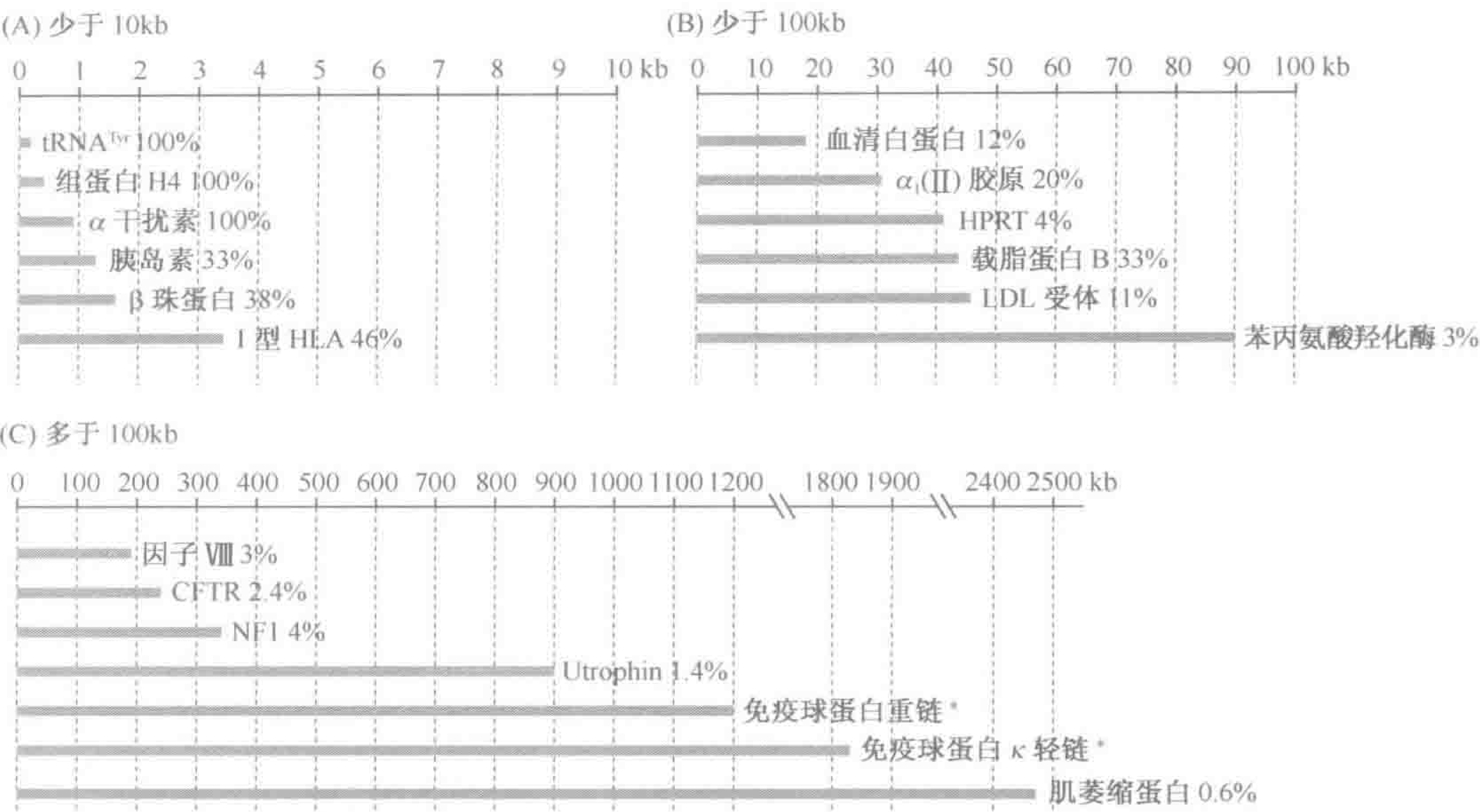


图 9.7　人类基因大小和外显子含量的巨大差异

外显子含量以标明基因长度的百分比（%）表示。**注**：基因长度和外显子含量的百分比通常是相反的关系。*强调标明 Ig 重链和轻链基因座的既定长度与生殖系的组成相一致（Ig 和 T 细胞受体基因有独特的组成，为在 B 或 T 淋巴细胞分别表达时，需要细胞特异的体细胞重排—见节 10.6）CFTR，囊性纤维化跨膜调节子；HPRT，次黄嘌呤磷酸核糖基转移酶；NF1，神经纤维瘤 I 型。

小与编码 DNA 片段间却有相反的关系（图 9.7）。这并非由于大基因的外显子比小基因的外显子小而引起的，虽然已知人类基因有很大的外显子（框 9.5），但其平均大小少于 200bp，外显子的大小相对地独立于基因的长度（表 9.6）。事实上，内含子的长度也有很大的不同，大基因趋向有很大的内含子（胶原 7 型和连接素基因则呈明显地例外-见表 9.6）。然而，长内含子的转录耗费时间和能量，而对高表达的基因，自然选择利于短的内含子（Castillo-Davis *et al.*,2002）。

表 9.5　具有连续编码序列的人类基因的举例

更详细的列表见http://exppc01.uni-muenster.de/expath/frames.htm

所有 37 个线粒体基因
许多 RNA 基因(特别是编码小 RNA 基因,例如,多数 tRNA 基因,还有一些大 RNA,例如 *XIST* RNA)
反基因(表 9.11)
干扰素
组蛋白基因
许多核糖核酸酶基因
热休克蛋白基因
许多 G 蛋白偶联受体基因
具有 HMG 框的一些基因(例如,*SRY*,许多 *SOX* 基因)
各种神经介质和激素受体基因,例如,多巴胺 D1 和 D5 受体。5-HT$_{1B}$色胺受体、血管紧张素Ⅱ1 型受体、甲酰多肽受体、舒缓激肽 B2 受体、α2 肾上腺素能受体。

框 9.5　人类基因组和人类基因的统计数字

基因组大小	～3200Mb
核基因组	～3200Mb
线粒体基因组	16.6kb
常染色质成分	约 2900～3000Mb
组成性异染色质	＞200Mb（表 9.2；图 2.15）
高度保守片段	＞100Mb（＞3%）
编码 DNA	＞50Mb（～1.5%）
其他（调节片段等）	＞100Mb（3%）
节段性复制 DNA	＞150Mb（＞5%）
非编码重复 DNA	＞50%基因组
基于转座子的重复	～1400Mb（～43%；见表 9.15）
基因数目	约 30 000～35 000
核基因组	约 30 000～35 000（节 9.1.3）
线粒体基因组	37（节 9.1.2）
每条染色体	平均～1400，但取决于染色体长度和类型（图 8.4）；在一550 带的染色体标本，每带～60
编码多肽基因	～30 000，但相当不确定
RNA 基因	～3000，但有的不确定（图 9.4）
假基因	～20 000
基因密度	～1/100kb 在核基因组，1/0.45kb 在线粒体基因组
基因大小（基因组范围）	平均＝27kb；但有很大的差异（图 9.7）
基因间距离	平均＝ca. 75kb 在核基因组
CpG 岛数目	～30 000（在基因组序列中除去非编码重复序列）
外显子数目	平均＝9，一般与基因长度有关，但变异很大
最大数	363（连接素基因内）
最小数	1（即无内含子-表 9.5）
外显子大小	相对没有长度的变化的内部外显子，平均＝122bp，但 3′的外显子有相当的长度（Zhang，1998）
最大的外显子	许多 kb 长，例如，*apoB* 基因（*APOB*）的外显子 26 长 7.6kb
最小的外显子	＜10bp
内含子大小	巨大的差异，与基因大小最直接相关（表 9.6）
最大的内含子	几百 kb，例如，人类 *WWOX* 基因的内含子 8 长～800kb
最小的内含子	几十 bp
mRNA 大小	平均大小约 2.6kb，但变化相当大（titin 基因 mRNA 长＞115kb!）
5′UTR	平均大约 0.2～0.3kb
3′UTR	平均大约 0.77kb，但可能估计不够，由于 3′UTR 长度的报道不完全
非编码 RNA 大小	极其不同，从约 21～22 核苷酸（microRNA）到许多 kb [例如 *XIST*（17kb）]
多肽大小	平均大约 500～550 氨基酸
最大的多肽	连接素：有 38 138 个密码子（但有显著的长度变化）
最小的多肽	几十个氨基酸，例如一些小激素等

表 9.6　人类基因外显子和内含子的平均大小

基因产物	基因大小(kb)	外显子数	外显子平均大小(bp)	内含子平均大小(bp)
$tRNA^{tyr}$	0.1	2	50	20
胰岛素	1.4	3	155	480
β 珠蛋白	1.6	3	150	490
HLA1 型	3.5	8	187	260
血清白蛋白	18	14	137	1100
胶原Ⅶ型	31	118	77	190
补体 C3	41	29	122	900
苯丙氨酸羟化酶	90	26	96	3500
因子Ⅷ	186	26	375	7100
CFTR(囊性纤维化)	250	27	227	9100
连接素	283	363	315	466
肌萎缩蛋白	2400	79	180	30 770

重复 DNA 含量的多样性

基因在非编码的内含子和旁侧序列内通常含有重复 DNA 成分。而且在编码 DNA 也有不同程度的重复 DNA 序列。串联重复的**微卫星序列**（microsatellite sequence）（短序列基序，节 9.4.3）是常见的并且可从统计学上简明反映一定的碱基成分的预期频率。编码已知的或推测的**蛋白质结构域**（protein domain）的串联重复序列也十分常见，并可能在功能上有用，如在某些情况下提供有用的生物学靶标。在有些情况下，重复序列间的序列同源性很高，在其他情况下，同源性则可能很低（表 9.7）。

表 9.7　大尺度基因内重复编码 DNA 的例子

基因产物	编码重复氨基酸的大小	拷贝数	拷贝间核苷酸序列同源性
内皮蛋白	10	59	中央 39 个重复高度同源
载脂蛋白？(a)	114＝kringle 4 样的重复[a]	37	高度同源：24 个重复在序列上相同
血纤维蛋白溶原酶	约 75～80	5	低度同源但有保守蛋白域(kringles[a])
胶原	18	57	低度同源但有保守的氨基酸基序$(Gly\text{-}X\text{-}Y)_6$
血清白蛋白	195	3	低度同源
富含辅氨酸蛋白基因	16～21	5	低度同源
弹性肌球蛋白 α-链	42	7	低度同源
免疫球蛋白 ε-链，C 区	108	4	低度同源
肌萎缩蛋白	109	24	低度同源

a kringle 是一富含半胱氨酸序列，它内含三个双硫桥并形成椒盐卷饼形的结构。

9.3.2　功能相似的基因偶尔簇集于人类基因组中，但更通常散在于不同的染色体上

正如在节 9.2 中所见到的，RNA 基因的某些家族是簇集的。至于编码多肽基因家族，编码相同的产物或在序列上非常紧密相关的基因常见于一个或多个簇，这些簇可散在于若干染色体上。然而，有些仅编码保守成分（结构域，重要的基序等）产物的功能

基因家族，常散在于基因组中。编码功能上相关但无明显序列同源产物的基因，则独特地散在于不同染色体。

功能相同的基因

极少的人类多肽已知是由两个或更多相同的基因拷贝编码的。通常，在一个基因簇内这些多肽是由最近复制的基因编码，例如，复制的 α 珠蛋白基因。此外，在不同染色体上有些基因很偶尔地编码相同的多肽，例子包括：

- 组蛋白基因。NHGR1 组蛋白序列数据库收录总共 86 个不同的组蛋白序列，它们分布在 10 个不同的染色体上，虽然有两大簇在 6p 上（http://genome.nhgri.nin.gov/histones/chrmap.shtml/）但某些亚家族成员是相同的，尽管是由不同染色体上的基因所编码。
- 泛素蛋白基因。76 个氨基酸的**泛素**（ubiquitin）是一很保守的蛋白质，它在蛋白质降解和细胞压力反应中起着重要作用。人的泛素基因可见分布于若干染色体上的不同基因座。有些还见于一系列串联全长编码重复序列，进行共转录（多顺反子的转录单位）。其他是单体（但与核糖体蛋白质基因融合，构成双顺反子的转录单位）（Nei *et al*.,2000 及节 9.3.3）。

功能相似的基因

大部分人类基因是不同基因家族的成员，其中每个基因紧密相关但是在序列上不相同。许多这样情况的基因是簇集的并由串联基因复制而引起，正像 α 珠蛋白和 β 珠蛋白基因簇的每个不同成员的情况（图 9.11）。但那些位于不同染色体上编码产物明显相关的基因一般关系不大，正如 α 珠蛋白和 β 珠蛋白基因那样。然而，就 *HOX* 同源框基因家族来说，该基因家族是由四条染色体的每条上的大约 10 个基因组成的基因簇所构成，在不同染色体上的每个基因可能比同一基因簇内的成员相互间更为相关（图 12.9）。除上述外，编码紧密相关的组织特异性异构体或亚细胞结构特异性同工酶的基因通常位于不同的染色体上（表 9.8）。

表 9.8 编码功能相关产物基因的分布

基因编码	组成	例子
相同的产物	经常簇集但也可以分散于不同染色体	11p 上两个 α 珠蛋白基因(图 9.11)；编码 rRNA 基因(图 10.2)；某些组蛋白亚家族(见 http://genome.nhgri.nih.gov/histones/chrmap.shtml)
组织特异蛋白异构体或同工酶	有时簇集；有时非同线的	簇集的胰腺和唾液腺淀粉酶基因(1p21)；在骨骼肌(1p)和心肌(15q)表达的非同线的 α 肌动蛋白基因
不同细胞成分的同工酶	一般非同线的	细胞质的(c)和线粒体的(m)各种酶的同工酶。例如，乙醛脱氢酶(c)-9q 和(m)-12q；胸腺嘧啶激酶(c)-17q 和(m)-16q
在同一代谢通路的酶	一般非同线的	编码类固醇发生的酶的基因；类固醇 11-羟化酶-8q；类固醇-17 羟化酶-10q，类固醇 21-羟化酶-6q
同一蛋白质的亚单位	一般非同线的	α 珠蛋白-16p 和 β 珠蛋白-11p；铁蛋白重链-11q 和铁蛋白轻链-22q
信号通路的相互作用成分	一般非同线的	JAK1-1p；STAT1-2q
配体加相关受体	一般非同线的	胰岛素-11p 和胰岛素受体-19p；干扰素 β-9p；干扰素 β 受体-21q

功能相关的基因

某些基因编码的产物可能在序列上并不相关，但功能上明显相关。这些产物可能是某些蛋白质或大分子结构的亚单位，相同代谢或发育通路的成分，或可能是相互特异结合的需要，正像配体及其相应的受体结合那样。几乎所有这样的情况，基因都不是簇集的而是通常位于不同的染色体上（见表 9.8 的某些例子）。

9.3.3 人类基因组中偶见重叠基因，基因内基因和多顺反子转录单位

双向的基因组成和部分重叠的基因

简单基因组的基因密度高（人类线粒体、大肠杆菌、酿酒酵母分别每 0.5kb 、 1kb 和 2kb 约有 1 个基因），并经常有部分重叠基因的例子。有时一常见的有义链可用不同的读框（图 9.3）。复杂有机体的基因很少是簇集的（在人类核基因组每 100kb 只有一个基因）且重叠基因也不常见。然而，偶尔也会发现很紧密的相邻基因。有些基因的 5′端被几百个核苷酸分离，并从互补链转录，像这样双向的基因组成常见于 DNA 修复基因，例如可提供基因对的共同调节的情况（Adachi and Lieber，2002）。

部分重叠的基因（partially overlapping gene）在哺乳动物的复杂核基因组中是很少见的，重叠基因一般由两个不同的 DNA 链转录。强有力的基因簇集发生在富含 GC 的亚染色体区，特别在高密度基因区经常显示一些重叠基因的例子。例如，在 6p21.3 HLA 复合体的Ⅲ类区，平均基因密度大约为 1/15kb 并已知若干含有重叠基因的例子（图 9.8A）。

基因内基因

小核仁 RNA(snoRNA) 基因是不常见的，大多数位于其他基因内，经常编码核糖体相关的蛋白质或核仁蛋白质。这种安排能维持使同等的核糖体的蛋白质产物和 RNA 成分共同（Tycowski *et al.*,1993）。除 snoRNA 基因外，某些另外的基因，包括一些编码蛋白的基因，位于大基因的内含子之中。明显的例子是：*NF1*（神经纤维病Ⅰ型）基因（三个小的内部基因由互补链转录，图 9.8B）、*F8*（凝血因子Ⅷ）基因（两个内部基因按反向转录，图 11.20）和 *RB1*（视网膜母细胞瘤易患性）基因（一个内部基因由互补链转录，图 9.19）。

多顺反子转录单位

多顺反子的（=多基因的）转录单位常见于细菌的简单基因组，也十分常见于秀丽新小杆线虫。简单的人类线粒体基因组（节 9.1.2）和主要的 rRNA 基因簇（图 10.2）提供基因组中两个多顺反子转录单位的例子。此外，已知在核基因组中存在编码多肽的**双顺反子转录单位**（bicistronic transcription units）的罕见例子：从一个基因开始转录，通过相邻的下游基因继续转录以产生一前体蛋白质并被切割为不同的蛋白质。

胰岛素的 A 和 B 链可认为是来自一双顺反子转录单位（图 1.23），但它们在功能上

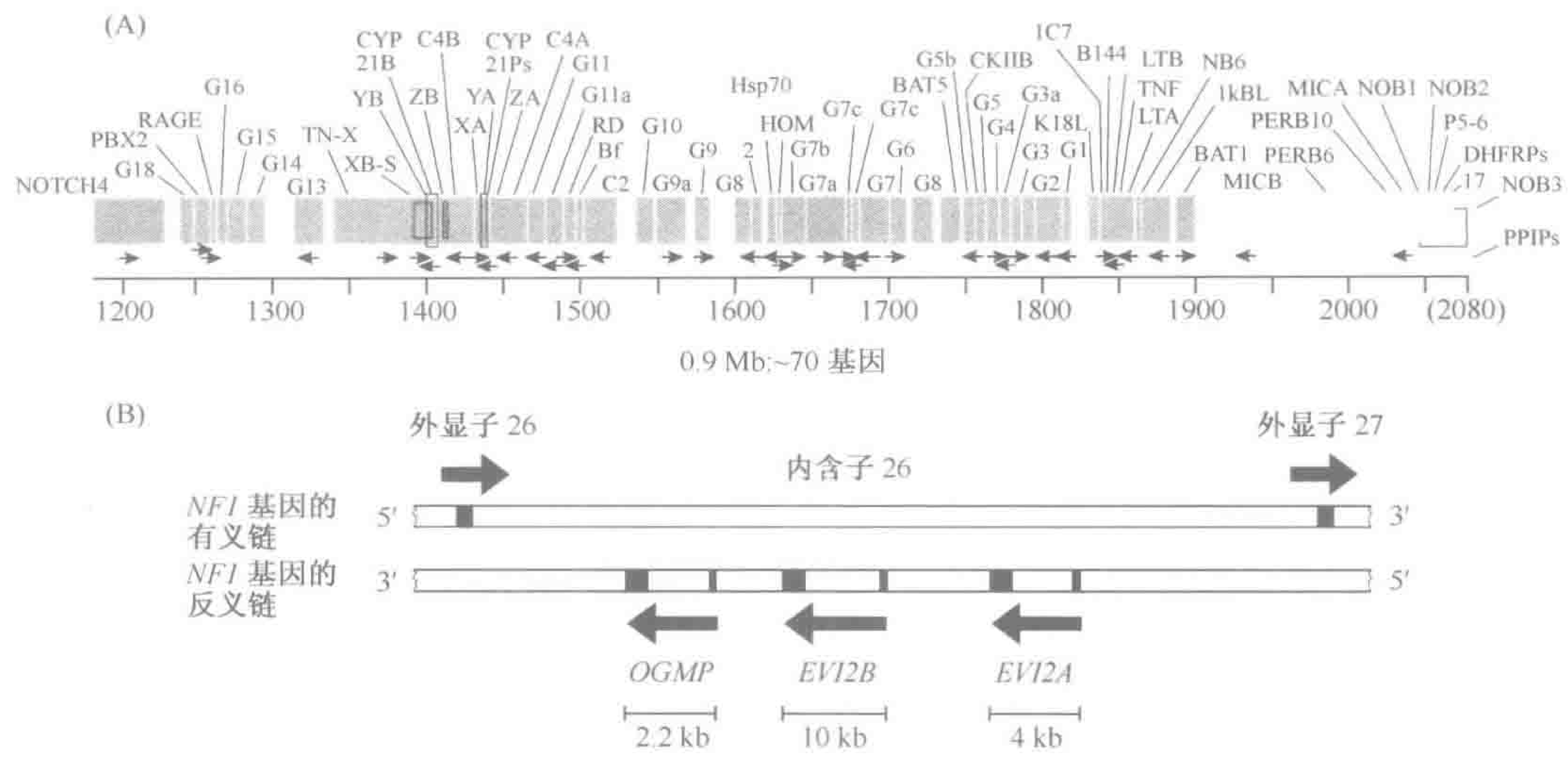


图 9.8　重叠基因和基因内基因

（A）**重叠基因**（overlapping gene）。在 HLA 复合体Ⅲ类区的基因，在某些情况下紧密包装和重叠。（B）**基因内基因**。神经纤维瘤Ⅰ型（*NF1*）基因的第 26 内含子内部含有三个两个外显子的基因，每个都由反链转录以用来转录 *NF1* 基因。基因是：*OGMP*，寡树枝细胞髓磷脂糖蛋白；*EVI2A* 和 *EVI2B*，鼠基因的人类同源基因，涉及白血病发生，位于向异病毒整合位点。

密切相关，然而，有时双顺反子转录单位产生功能上不同的蛋白质。例如，*UBA52* 和 *UBA80* 基因分别生产泛素和核糖体蛋白质，S27a 或 L40。其他泛素基因由多顺反子转录单位转录为串联完全编码重复序列组成（Nei *et al.*,2000）。泛素基因没有内含子，但在其他的双顺反子转录单位中，剪接需要将一个基因外显子转录物与下游基因外显子转录物连接起来。*SNURF-SNRPN* 转录单位提供了由不同外显子编码两个多肽的例子（Gray *et al.*,1999），而且也合成非翻译 RNA 转录物，它们是父性印记（图 10.24）。

9.3.4　编码多肽基因家族可按家族成员的序列相关性的程度和范围分类

大部分活跃表达的，编码非编码 RNA 和多肽的人类基因是 **DNA 序列家族**（DNA sequence family）成员，该家族显示高度的序列相似性。然而，家族成员在共享序列的范围和组成上都有很大的不同。许多家族成员可能是无功能的（假基因和基因片段-见下文）并快速积累序列差异导致明显的序列趋异。

典型的基因家族

典型的基因家族成员在基因长度的大部分，或至少在编码 DNA 的成分上显示高度的序列同源性。例子包括组蛋白基因家族（组蛋白是很保守的，同时其亚家族成员实质上是相同的）和 α 珠蛋白基因及 β 珠蛋白基因家族（每个家族的成员显示高度的序列相似性）。

编码具有大而高度保守结构域产物的基因家族

某些基因家族的成员在基因的特别强的保守区内具有特别显著的同源性。在不同基因的编码序列的保守部分之间的相应序列其相似性可能十分低。这样的家族通常编码在早期发育中起重要作用的转录因子，而且保守序列编码与选择靶基因的 DNA 特异结合所需要的一蛋白质结构域（表 9.9）。

表 9.9　具有编码高度保守结构域序列基序的人类基因的例子

基因家族	基因数目	序列基序/结构域
同源框基因	38 *HOX* 基因(图 12.9)加 214 孤儿同源框基因	同源框(homebox)特化一60 氨基酸的同源结构域(homeodomain)，已确定了许多不同的亚类
PAX 基因	9	成对框(paired box)编码～128 氨基酸的成对结构域(paired domain)，*PAX* 基因常有另一类型同源结构域，称为成对性结构域
SOX 基因	18	SRY 样 HMG 框，编码 69 氨基酸的结构域
TBX 基因	18	T-Box，编码 170 氨基酸的结构域
叉头结构域基因	49	叉头结构域(forkhead domain)，长～110 氨基酸
POU 结构域基因	24	POU 结构域(POU domain)，长～150 氨基酸

编码具有短而保守的氨基酸基序产物的基因家族

某些基因家族的成员在 DNA 序列水平不是明显相关，不过编码基因产物具有常见的一般功能并有很短的保守序列基序，诸如，DEAD 框，此序列为 Asp-Glu-Ala-Asp (DEAD 以氨基酸密码子的一个字母表示)，或 WD 重复（tryptophan-aspartate，色氨酸-天冬氨酸）（图 9.9）。

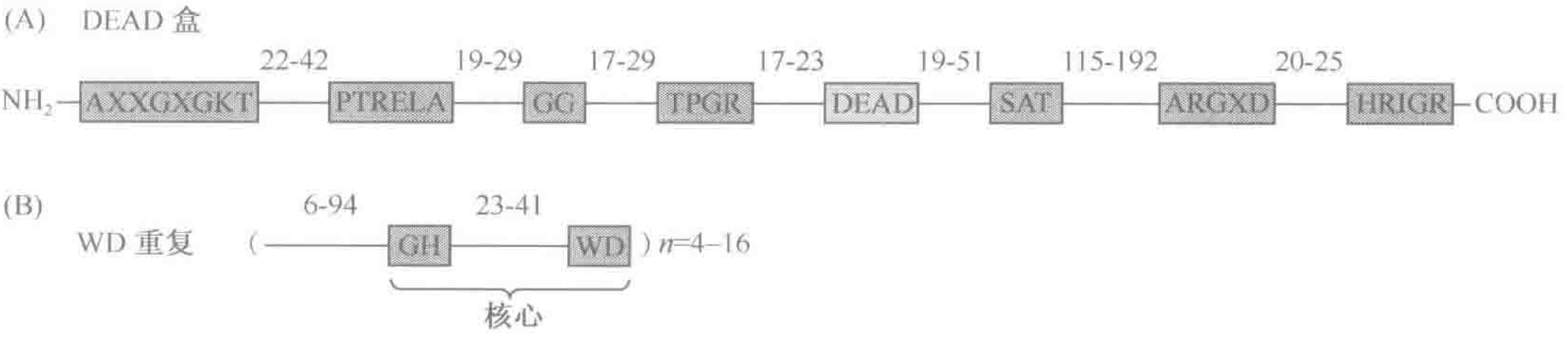


图 9.9　某些基因家族由具有很短保守的氨基酸基序的功能相关基因产物所限定

(A) DEAD **框家族的基序**。这个基因家族编码产物与涉及选择性 RNA 次级结构的细胞内加工有关，诸如翻译起始和剪接。八个很保守的氨基酸基序就是明显的证据，包括 DEAD 框（Asp-Glu-Ala-Asp）。数字表示常见的氨基酸序列之间的大小范围（见 Schmid 和 Linder，1992）。X＝任一氨基酸，见氨基酸密码的单字母表。(B) WD**重复家族**。此基因家族编码产物与各种调节功能有关，诸如调节细胞分裂、转录、跨膜信号、mRNA 修饰等。基因产物以 4～16 串联 WD 重复组成约 44～60 氨基酸为特征，每个含固定长度的核心序列，以 GH(Gly-His) 开始和以 WD(Trp-Asp) 二肽为终止，但有不同长度的序列加在前面（Smith *et al.*，1999）。

基因超家族

在进化中基因超家族成员之间的关系比一个典型的或含保守结构域/基序的基因家族成员之间更远。它们所编码的产物，在一般意义上功能相关，且表明在大片段序列上只有很弱的同源性，也无很重要的保守氨基酸基序。而都有一些一般常见的结构特征和一般相关功能的证据。举例如下：

- **免疫球蛋白超家族**（immunoglobulin superfamily）（图 9.10）——一个很大的家族，包含有免疫球蛋白（Ig）基因，T 细胞受体基因，HLA 基因和许多其他基因，基因编码产物在序列水平又相当地不同，但都具有免疫系统功能并有 Ig 样的结构域；

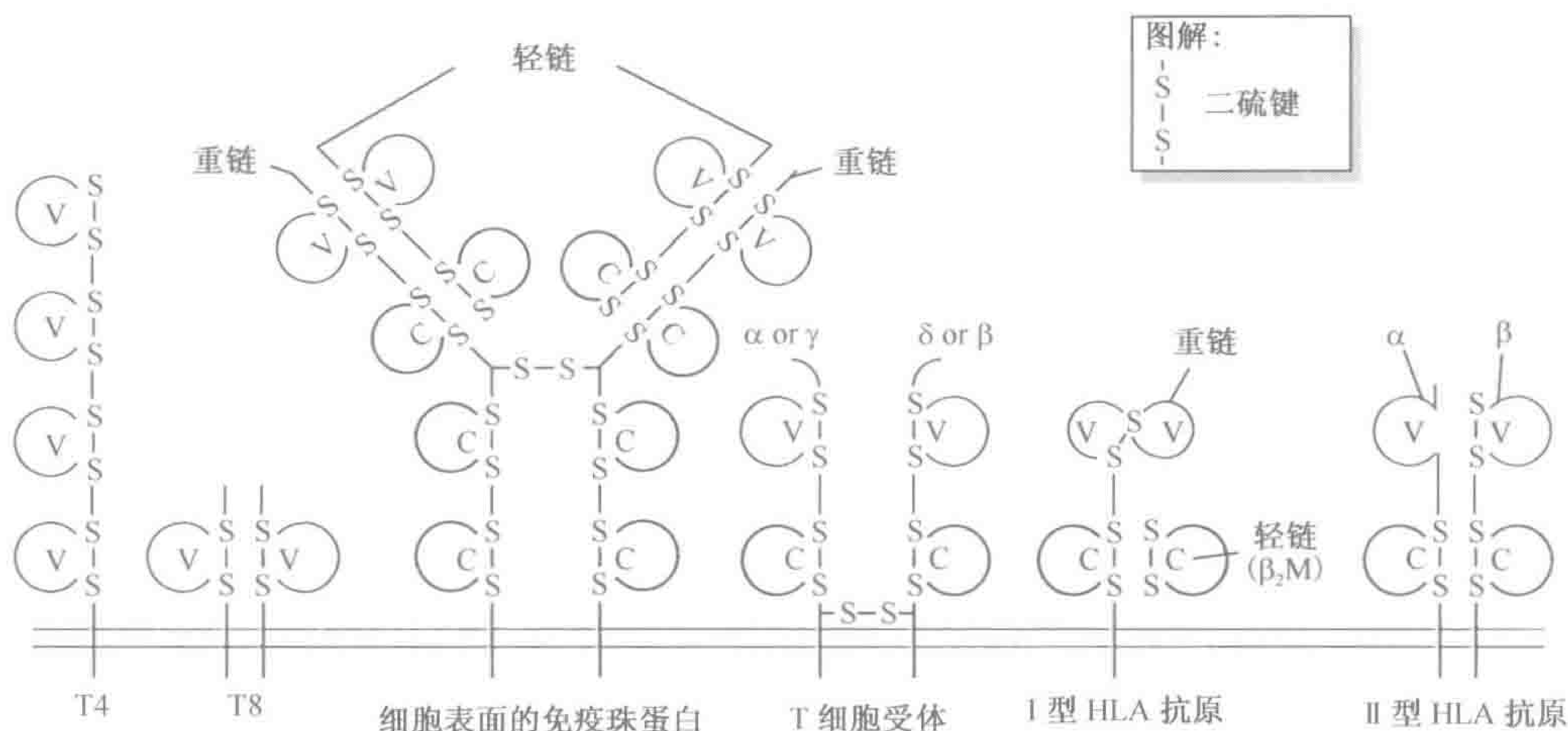


图 9.10　Ig 超家族成员是具有相似类型功能结构域的表面蛋白质

图示很大的 Ig 超家族的一些例子。很多成员是由在 N 端的细胞外可变域（V）和在 C 端的恒定域（C）组成的双体。Ⅰ型 HLA 抗原轻链，β_2小珠蛋白有一简单的恒定域但并不横跨膜。它与含有两个可变域和一个恒定域的膜重链有关，但其整体结构相似于Ⅱ型 HLA 抗原。

- **珠蛋白超家族**（globin superfamily）——一个小家族，不仅包含在血液中运输和贮存氧的 α 珠蛋白和 β 珠蛋白基因家族的成员，而且还包括分别编码肌肉和脑的珠蛋白，即肌红蛋白和脑红蛋白的基因（图 12.4）。
- **G 蛋白偶联受体超家族**（G protein-coupled receptor superfamily）——一个很大而多样的受体家族，通过与细胞内 G 蛋白相互作用，该受体中介配体诱导细胞内外环境之间的信号。它们共有一常见的 7 个 α 螺旋跨膜节段结构，但其序列彼此间的相似性却是特别地低（<40%）。

9.3.5　人类基因家族中基因可以组成小簇，广泛分散或两者兼有

有证据表明人类基因家族可分为紧密的基因家族和分散于某些不同染色体位置上的基因家族。然而，因为某些基因家族由位于不同的染色体位置上的多基因簇组成（表 9.10），而其他的诸如组蛋白基因家族（http://genome.nhgri.nih.gov/histones/chrmap.shtml）可能由一个或两个大基因簇为主，但也有某些分散的为**孤独基因**（orphan

gene）家族，所以这样分类多少有些随意性。

表 9.10　簇集的和散在的多基因家族的例子

家族	拷贝数	组成	染色体的位置
(A)簇集的基因家族			
单簇集基因家族			
生长激素基因簇	5	簇集于 67kb 内；一个传统的假基因	17q22-24
α 珠蛋白基因簇	7	簇集超过～50kb(图 9.11)	16p13.3
Ⅰ型 HLA 重链基因	～20	簇集超过 2Mb(图 9.12)	6p21.3
多簇集基因家族			
HOX 基因	38	组成四个簇于染色体 2p、7、12、17 上(图 12.9)	
组蛋白基因家族	61	中等大小的簇在几个位置；两个大簇位于 6 号染色体	许多
嗅觉受体基因家族	>900	大约 25 个大簇分散于全基因组	许多
(B)散在的基因家族			
丙酮酸脱氢酶	2	含有一个内含子的基因和一个睾丸表达的反基因	Xp22;4q22-q23
醛缩酶	5	在 5 个不同染色体上有 3 个功能基因和 2 个假基因	许多
PAX	9	9 个全部都是功能基因	许多
NF1(神经纤维瘤Ⅰ型)	>12	一个功能基因在 17q；其他为缺损的非加工的 DNA 拷贝(图 9.13)	许多，大多为臂间着丝粒
铁蛋白重链	>15	一个功能基因在染色体 11；大多为加工的假基因	许多

以单个簇组成的基因家族

在各个基因簇内的基因通过**串联基因复制**（tandem gene duplication）事件而产生（图 12.3）。不同组成的证据如下：

▶ 串联基因组成。这些基因在序列和功能上彼此相关，虽然一些家族成员可能是无功能的。编码多肽基因的例子很少（多泛素基因是一个明显的例子），但某些 RNA 基因家族（rRNA，U2 snRNA）表明为这样的组成。

▶ 紧密成簇。有的基因并非完全串联重复，而是紧密成簇，并可由一单个基因座控制区来调节（图 9.11）。α 珠蛋白和 β 珠蛋白基因簇集的例子。每个基因通常表现彼此间序列和功能的高度一致性，但许多家族成员可能是假基因（节 9.3.6）。

▶ 复合基因簇。然而，在另外簇集的基因家族中，簇内的基因间的物理关系可能并不紧密而一簇相关的基因内也可能含有在序列和功能上无关的基因，构成**复合基因簇**（compound gene cluster）。例如，6p21.3 上的 HLA 复合体，由编码 HLA 抗原Ⅰ型和Ⅱ型的基因家族和各种血清补体基因所控制，但每个家族成员可由功能上无关的基因分开，如类固醇 21 羟化酶基因家族的成员等。

由多基因簇组成的基因家族

某些基因家族由多基因簇组成。偶尔，这些簇可能在同一染色体上紧密相关，是近期（进化中）复制的结果。例如，反转簇含有脊椎肌肉萎缩相关的 *SMN1* 和 *SMN2* 基因（Frogier *et al*.,2002）。然而，它们通常分布在两个或更多的染色体位置上。不同的组成

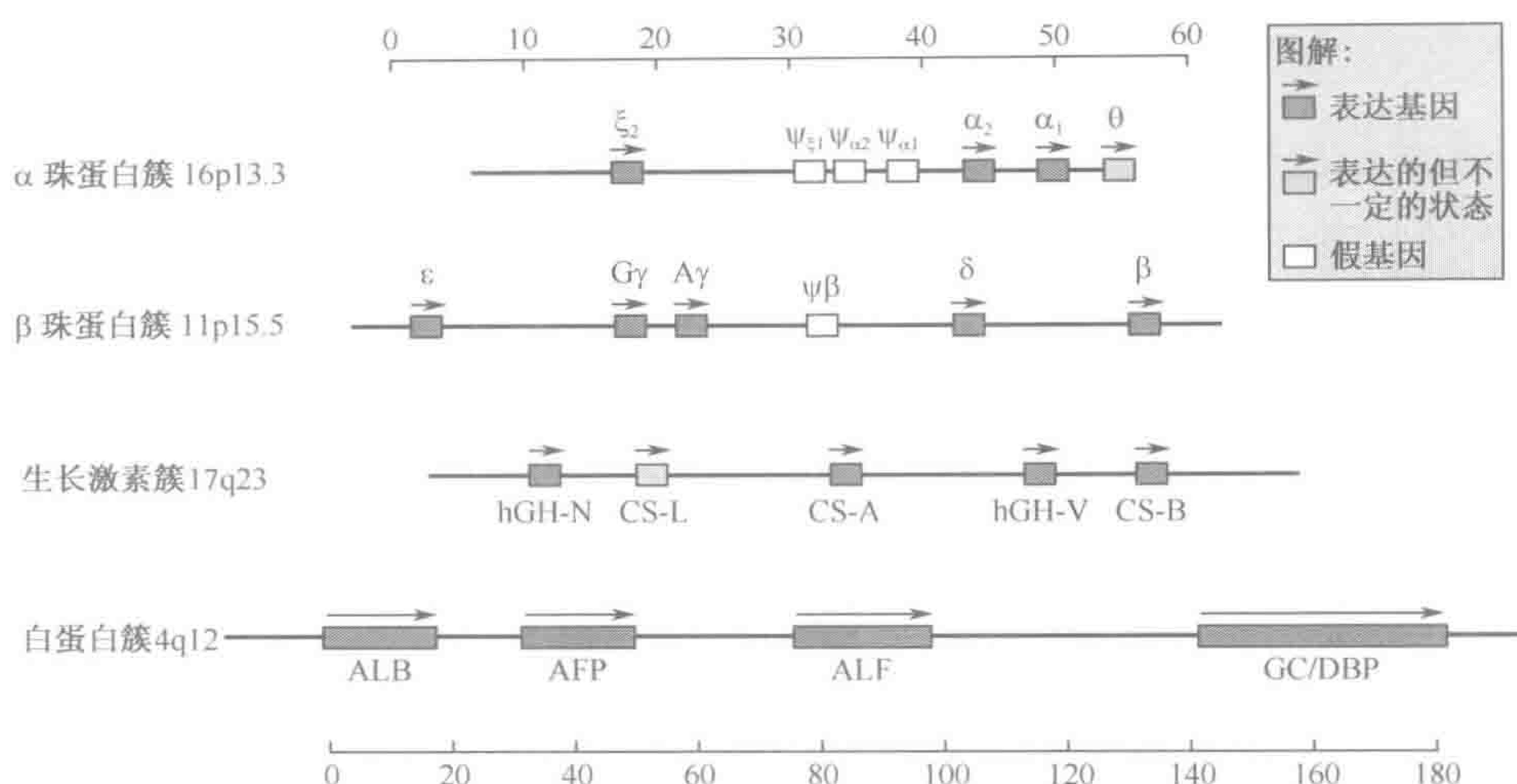


图 9.11 人类簇集的基因家族的例子

一个簇中的基因在序列上紧密相关，且典型地由相同的链转录。θ 珠蛋白和 *CS-L* 基因的功能状态尚不清楚。顶部（珠蛋白和生长激素簇）和底部（白蛋白簇）的大小以 kb 计。

都有证据。某些家族表明在不同簇的基因之间有较高的相似性；而某些就少些。一个最明显的例子是：嗅觉受体基因家族，它编码组成多样的受体，使我们能识别上千种不同的气味。这个>900 成员的家族是由大的基因簇组成，这些基因簇位于 25 个以上的不同的染色体位置，即除 20 号和 Y 染色体以外的所有染色体（Glusman *et al*.,2001）。

序列同源性一般是簇内大于簇间（例如，比较 16p 上的 α 珠蛋白簇和 11p 上的 β 珠蛋白簇的成员，图 12.4），但偶尔由于强有力的功能性选择，在不同簇内的基因比在单一簇内彼此间更加相关，如 *HOX* 基因的情况（图 12.9）。

散在的基因家族

某些家族成员分散在两个或更多不同的染色体位置。在不同位置上的基因除非最近发生基因复制或曾因相当大的选择压力以维持序列的保守性，一般在序列上是十分趋异的。家族成员可由以下情况组成：

- 不同的基因组。原始的线粒体基因组可能来自一种有氧细菌。它们相继转运许多原始细菌的基因到核基因组。结果，核基因组含有复制的基因，即编码细胞质特异的和线粒体特异的一些酶和其他关键代谢产物的异构体的基因（见表 9.8 的某些例子）。
- 古基因/基因组复制事件。正如 *PAX* 基因家族的情况，这类典型的家族只有少数成员而且在进化的长河中经历基因复制和/或基因组复制结合的事件，已显示了进化。通常所有或许多的家族成员是有功能的，而且在基因产物间显著的序列同源性可能受到重要的决定性结构域的限制，例如 *PAX* 基因产物的成对结构域。
- 反转座的事件。某些基因家族通过进化的过程具有较近期的扩张，由此，从一个或少数功能基因转录的 RNA，由细胞的反转录酶转化为天然的 cDNA，然后它随处整合于染色体上。大多这样的拷贝是非功能性的，但某些基因家族有一功能性含内含子的基因和一功能性加工的基因拷贝（下一节）。

9.3.6 假基因、截短基因拷贝和基因片段常见于多基因家族

编码多肽基因的家族（和 RNA 基因）的常见特征是一个功能基因（或至少是其编码序列）的全部主要序列或它的部分序列有缺损的拷贝（**假基因**，pseudogene），例如，缺少 5′或 3′端的**截短拷贝**（truncated copy）或**内部片段**（internal fragment）。在某些情况下，是单个外显子。已发现许多不同的类型。下述的例子正表明在不同类型的基因家族中发现的有缺陷基因拷贝的类型。

基因簇中非加工的假基因

每个基因簇经常都有缺陷的基因拷贝，它们是由串联**基因复制**（gene duplication）在基因组 DNA 水平拷贝的。这些拷贝含有与功能基因（**非加工的假基因**，nonprocessed pseudogene）的外显子、内含子和启动子区相应的序列，但它们通常由与外显子相应的序列中存在不适当的末端密码子来识别缺损的拷贝，典型的例子可见于 α 珠蛋白和 β 珠蛋白基因簇（图 9.11）。

基因簇中截短基因和内部基因片段

HLA Ⅰ 型基因家族位于 6p21.3，是基因簇的一个典型例子，它由非加工的假基因、截短基因和基因片段所构成。虽然，HLA Ⅰ 型基因的数目在不同的 6 号染色体上有所不同，全面分析其中之一表明在约 2Mb 上簇居 17 个家族成员：6 个表达基因，4 个传统的全长假基因，5 个截短的基因拷贝和两个小的内部基因片段（Geraghty *et al*.，1992；图 9.12）。该家族起源于串联基因复制和由不等交换或姐妹染色单体不等交换产生的片段基因拷贝（节 11.3.2）。

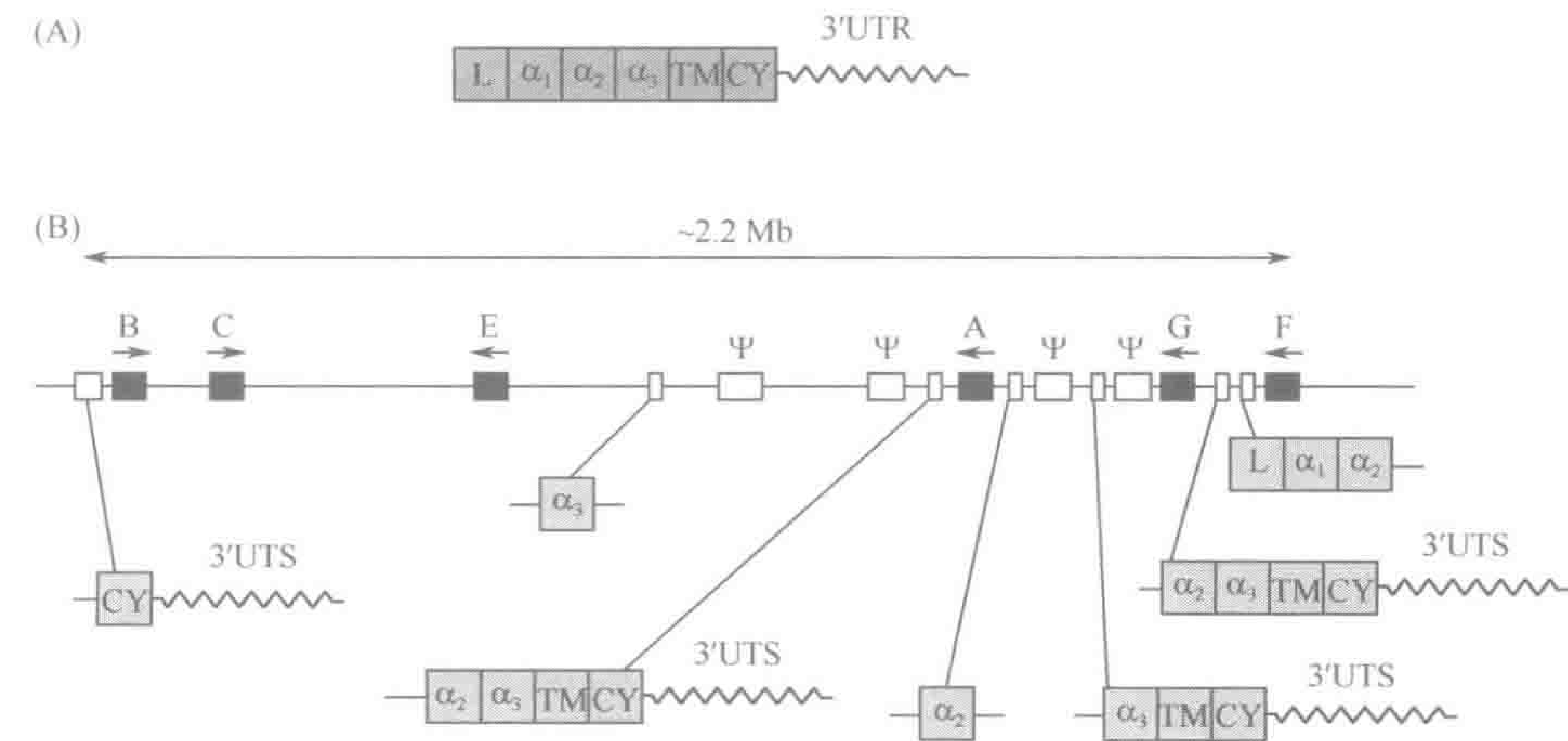


图 9.12 簇集基因家族经常含有非加工假基因和截短基因或基因片段：例如 HLA Ⅰ 型基因家族
（A）**HLA Ⅰ 型重链 mRNA 的结构**。其全长 mRNA 含有编码多肽序列；方块表示如下的不同的结构域：L，引导序列，α_1，α_2，α_3，细胞内结构域；TM，跨膜序列；CY，细胞质的尾和 3′-非翻译序列（3′-UTS）。这三个细胞外结构域 α_1-α_3，每个都由单个外显子编码，而很小的 5′UTS 没有表示。（B）**HLA Ⅰ 型重链基因簇**。该簇位于 6p21.3，约由 20 个基因组成。它们包括 6 个表达基因（黑色），4 个全长非加工假基因（Ψ），和不同的部分基因拷贝（空框）。后者有些是在 5′端被截短（例如，邻近 HLA-B 的那个），某些是在 3′端截短（例如，邻近 HLA-F 的那个），有些含有单个外显子（例如，邻近 HLA-E 的那个）。

散在的基因家族中非加工的假基因

NF1（神经纤维瘤Ⅰ型）和 *PKD1*（成人多囊肾病）基因相关的序列是两个明显的例子。这两个基因分别位于 17q11.2［靠近着丝粒（臂间着丝粒的）］和 16p13.3［靠近端粒（亚端粒的）］。人类的臂间着丝粒区是典型的由在进化过程中位于若干染色体上最近的拷贝序列所组成。亚端粒区也是较不稳定的，并易于复制（Eichler，2001；Mefford and Trask，2002）。它们在本质上参与占人类基因组 150Mb 以上的灵长类特异的**节段性复制**（segmental duplication）的形成，虽然这种不稳定效应似是特殊的部分染色体（Bailey *et al.*，2002；节 12.2.5）。

NF1 基因的情况，至少 11 个非加工的假基因/基因片段拷贝（含有相似于 *NF1* 内含子和外显子的序列）分布在 7 个不同的染色体上，9 个位于臂间着丝粒区（Regnier *et al.*，图 9.13）。*PKD1* 基因有 46 个外显子，全长 50kb。一个截短的 5′基因拷贝组分约占该基因的 70%（外显子 1～34 加上间隔的内含子），曾至少实在地复制过三次并插至 16p13.1 较近端的位置（欧洲多囊肾病协作组，1994）。

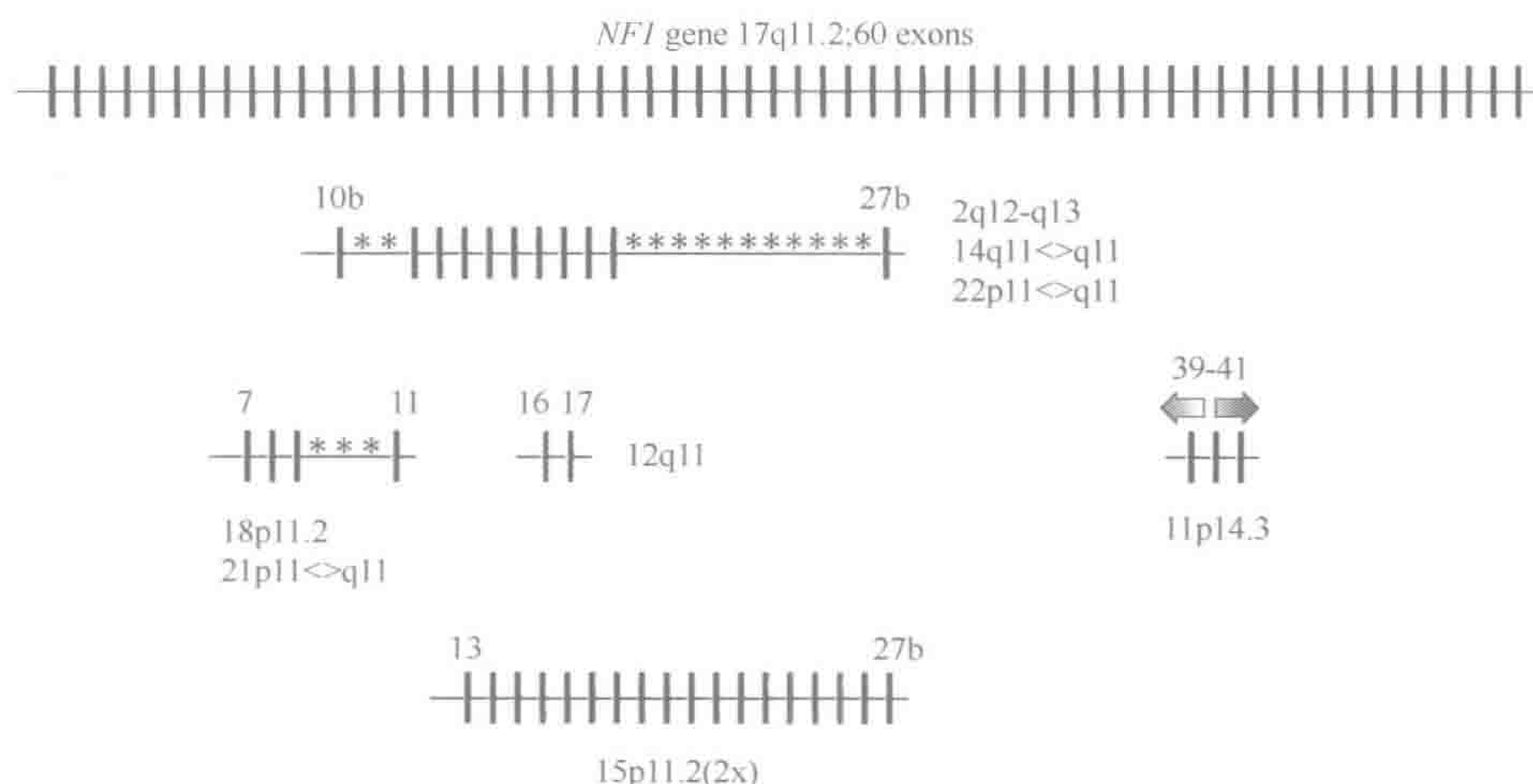


图 9.13 散在非加工的假基因起源自臂间着丝粒的 *NF1*（神经纤维瘤Ⅰ型）基因

细的垂直框表示外显子，为方便起见，*NF1* 基因的内含子以等长表示。虽然 *NF1* 基因有 60 个外显子，外显子的号码从 1～49 与其某些相邻外显子都以相同的数目设定，而以字母区别（例如外显子 10a，10b 和 10c）。*NF1* 基因内高度同源假基因拷贝可见于 8 个或更多其他的基因组位点，多数在臂间着丝粒区。在每种情况下，假基因仅由全长基因的一部分拷贝所组成，具有一些外显子及其之间的内含子。假基因的重排是显而易见的。有些重排已引起外显子和内含子的缺失（由 * 号表示）。其一个假基因已经倒位，因此 39 外显子拷贝相对于相邻的外显子拷贝反转了。数据由 Wales 大学医学院 Nick Thomas 博士和 Meena Upadhyaya 博士提供。

散在的编码多肽基因家族中加工的假基因

散在的基因家族经常有缺损的基因拷贝，它们含有与一个功能基因的外显子相应的序列（但不是内含子），同时还常在一末端含有一寡（dA）/（dT）序列。这样**加工假基因**（processed pseudogene）是由反转座子在 cDNA 水平拷贝的（节 9.5.1）。细胞反转

录酶转录 mRNA 为天然的 cDNA，然后它能整合进染色体 DNA（图 9.14），此过程最大的可能是通过 LINE1 转座装置来协助的（节 9.5.2）。加工的假基因能大大增殖。例如，在细胞质的核糖体上有 79 个蛋白质和 95 个核糖体蛋白质基因的家族（16 个是复制的），但在核基因组中却鉴定了数目大得惊人的 2090 个这种类型的加工假基因（Zhang，2002）。

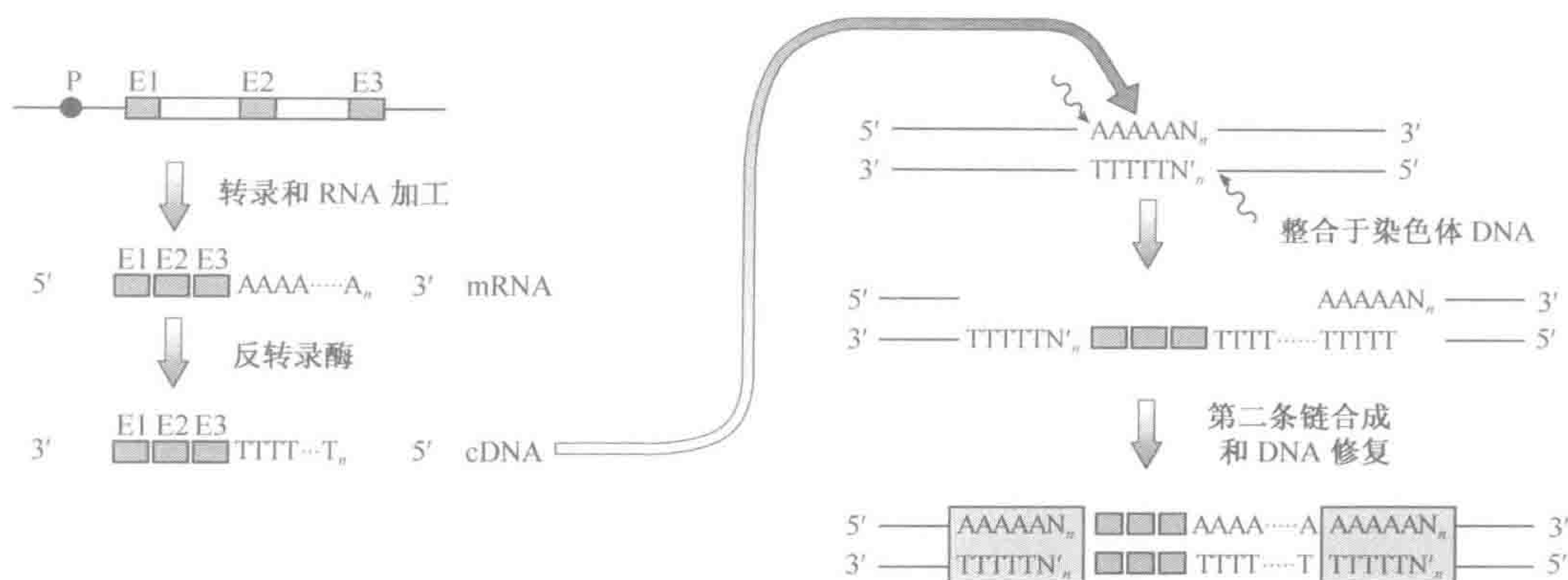


图 9.14　由 RNA 转录物反转录产生的加工假基因和反基因

LINE1 重复可提供反转录酶功能。图中所示整合模式只是几种可能性之一。这里设想整合是在富含 A 的序列链的断开处（以卷曲箭头指示），但由 LINE1 内切核酸酶协助（节 9.5.2）。如果此富含 A 的序列包括在 5′突出端，它能与 cDNA 的 poly(T) 远端进行杂交，促使第二链的合成。由于在整合过程中链的断裂，插入序列将会由短的直接重复（框内的序列）排列于两侧。E1-E3 是外显子。P=启动子。转座拷贝不携带启动子，因此它们在正常情况下不表达并将获得缺失突变（**加工的假基因**，processed pseudogene）。然而，某些编码多肽基因的拷贝是有功能的（**反基因**，retrogene），由于它们整合于相邻的功能启动子的位点，且其保守功能受选择压力的支配（表 9.11）。

典型的加工假基因不表达（因为它们缺少启动子序列），但已知某些例子是加工基因的表达。这里天然的 cDNA 曾碰巧整合进一染色体的 DNA 位点，偶尔靠近一个启动子，它能驱动加工基因拷贝的表达。选择压力可能确保这个加工基因拷贝的继续表达，这样的拷贝则称为**反基因**（retrogene）。许多没有内含子的反基因已知具有睾丸特异表达模式，而且它们常是一个含有内含子 X 连锁基因的常染色体同源物（表 9.11）。这里选择压力可能是男性减数分裂过程中表达所必需的，即当常染色体的基因转录而 X 和 Y 两个染色体的基因却是沉默的时候，并被浓缩形成异染色质。然而，某些功能反基因是非 X 连锁基因的拷贝，如像 *SNAIL1* 这个发育调节基因的拷贝（Locascio *et al.*，2002）。

编码 RNA 基因家族中加工的假基因

虽然某些散在的编码多肽基因家族的大小证明以反转座子作为产生加工基因拷贝的一种机制是成功的，但真正成功的（按其高拷贝数目）反转座是由 RNA 聚合酶Ⅲ转录物执行的。例如，Alu 重复家族（节 9.5.3）产生于编码 SRP RNA 的基因（也称为 7SL RNA），信号识别颗粒的一种成分，拷贝加工假基因像这样的由 RNA 聚合酶Ⅲ转录的基因，经常含有一内部的启动子（图 10.4），以促进新转座拷贝在基因组允许的区域内表达。

表 9.11 人类无内含子反基因及其父性含有内含子同源体的例子

（更多信息：见http://exppc01.uni-muenster.de/expath/frames.htm）

反基因	含内含子的同源物	产物
在 4q13 的 *GK2*	在 Xp21.3 的 *GK1*	甘油激酶
在 4q22-q23 的 *PDHA2*	在 Xp22 的 *PDHA1*	丙酮酸脱氢酶
在 6p12.3 的 *PGK2*	在 Xq13 的 *PGK*	磷酸甘油激酶
在 9p13.3 的 *TAF1L*	在 Xq13.1 的 *TAF1*	TATA 框结合蛋白相关因子，250kDA
在 1p34.2 的 *MYCL1*	在 Xq22-q23 的 *MYCL2*	v-myc 癌基因同源体
在 10q23.3 的 *GLUD1*	在 Xq25 的 *GLUD2*	谷氨酸脱氢酶
在 2q33-q37 的 *SNAIL1L1*	在 20q13 的 *SNAIL1*	蜗牛相关的发育调节子

9.3.7 人类蛋白质组分类虽已开始，但许多人类蛋白质的精确功能尚未确定

人类基因组的测序提供了预测**人类蛋白质组**（human proteome）的有价值信息。对于许多基因，蛋白质功能早先曾有所设定，但分析大量的新基因使早先的分类扩展了。各种数据库已致力于记录序列特征，这些特征为多蛋白质所共有并表现出共同的或相关的功能（虽然不是所有的蛋白质都能被归入已有的类目，因为某些蛋白质看来不像是与其他蛋白质共用序列）。常用的数据库包括由欧洲生物信息学研究所建立的 InterPro 数据库和在 Wellcome Trust Sanger 研究所建立的 Pfam 数据库（见进一步阅读）。类目包括：

▶ **蛋白质家族**（protein family）（根据一般功能的相似性，表 9.12）；

表 9.12 人类蛋白质组中最重要的 12 个蛋白质家族

资料获自 2003 年 1 月欧洲生物信息学研究所的 InterPro 数据库（http://www.ebi.ac.uk/proteome/）

InterPro reference	蛋白质家族名称	匹配的蛋白质
IPR000276	视紫红样 G 蛋白偶联受体	826
IPR000719	蛋白质激酶	688
IPR001909	KRAB 框（Kruppel 相关框）	314
IPR001806	RasGTP 酶超家族	192
IPR005821	离子运输蛋白	149
IPR000387	酪氨酸特异性蛋白磷酸酶和双特异性蛋白质磷酸酶	139
IPR001254	丝氨酸蛋白酶，胰蛋白酶家族	128
IPR000379	酯酶/脂酶/硫酯酶，活性位点	112
IPR007114	主要促进子超家族（MFS）	100
IPR001993	线粒体底物载体	86
IPR001664	中间纤维蛋白	85
IPR001128	细胞色素 P450	84

▶ **蛋白质结构域**（protein domain）（表 9.13）。大量的锌指证明它们在许多 DNA-蛋白质相互作用中的重要性；

表 9.13　人类蛋白组中最重要的 15 个蛋白质结构域

资料获自 2003 年 1 月 InterPro 数据库（http://www.ebi.ac.uk./proteome/）

InterPro reference	蛋白质结构域名称	在蛋白质组中总数
IPR007087	锌指，C2H2 类	28 654
IPR002126	钙连蛋白	4131
IPR006209	表皮生长因子(EGF)样结构域	3107
IPR003006	免疫球蛋白/主要组织相容性复合体	2387
IPR002048	钙结合 EF 手	1885
IPR001452	SH3 结构域	1815
IPR003961	纤连蛋白Ⅲ型	1812
IPR000504	RNA 结合区 RNP-1(RNA 识别基序)	1783
IPR001356	同源框	1435
IPR002965	富含脯氨酸延伸蛋白	1229
IPR001478	PDZ/DHR/GLGF 结构域	1143
IPR001841	锌指环	1132
IPR001849	类 Plekstrin	1061
IPR000210	BTB/POZ 结构域	494
IPR005225	小 GTP 结合蛋白质结构域	189

▶ **蛋白质重复**（protein repeat）。最常见的有 G-蛋白 βWD-40 重复（～400 蛋白质配对）和锚蛋白重复（>260 蛋白质配对）。

功能分类也已开始，同时**基因实体学协作组**［Gene Ontology(GO) consortium］依据蛋白质在细胞成分中的作用，其分子功能和它参与细胞的整体生物学过程已定义了功能分类类目（节 8.3.6）。当然，这是一发展过程：许多人类基因的功能仍需应用各种手段进行确定。依其分子功能和生物学过程进行了人类蛋白质的初期分类，列表于图 9.15。

图 9.15　根据它们的分子功能和参与的生物学过程对人类和小鼠蛋白质的分类

（A）**分子功能**（molecular function）（B）**生物学过程**（biological process）在较大的实体学内，基因实体学（GO）术语大约分为 12 类。灰棒：小鼠蛋白质；白棒：人类蛋白质。获自 Nature Publishing Group 的允许，修改自小鼠基因组测序协作组（2002）。Nature 420，520～562。

9.4 串联重复非编码DNA

高度重复非编码DNA经常存在一个序列的串联重复排列（或方块），序列可能是简单的（1～10核苷酸），或中等复杂的（几十到几百个核苷酸）。每一排列能发生在几个或许多不同的染色体部位。根据排列大小可分为三个主要亚类：卫星DNA，小卫星DNA和微卫星DNA（表9.14）。当卫星DNA绝大多数为小卫星DNA时，它在转录上是失活的，但相当比例的微卫星DNA（虽然是很小的）则位于编码DNA内。

表9.14 人类串联重复DNA的主要类型

类型	重复单位大小/bp	主要染色体的位置；转录状态
卫星DNA（序列常在100kb～几Mb范围内）	5～171	特别在着丝粒；不转录
α(alphoid DNA)	171	所有染色体的着丝粒的异染色质区
β(Sau3A家族)	68	1、9、13、14、15、21、22和Y的着丝粒的异染色质
卫星1(富含AT)	25～48	大多染色体的着丝粒的异染色质和其他异染色质区
卫星2和3	5	大多数，可能是全部的染色体
小卫星DNA(序列常在0.1～20kb范围内)	9～64	位于或靠近所有染色体的端粒，绝大多数不转录
端粒家族	6	所有的端粒
高可变家族	9～64	所有的染色体，常靠近端粒
微卫星DNA(=**简单序列重复**，SSR)（典型的序列<100bp)	12	分散于全部所有的染色体；某些很小的排列的简单序列

9.4.1 卫星DNA由很长的串联重复排列组成并能用密度梯度离心从一堆DNA中分开

人类卫星DNA由很大的串联重复DNA排列组成。重复单位可以是一个简单的序列（仅几个核苷酸长）或一中等复杂序列（表9.14；Singer，1982）。卫星DNA构成基因组的大多数异染色质区，值得注意的是常见于着丝粒的附近（**臂间着丝粒异染色质区**，pericentrometic heterochromatin）。当重复单位很短时，重复单位的碱基成分和卫星DNA全部的碱基成分，在本质上可与总体基因组的DNA碱基成分分开。结果，可用buoyant密度梯度离心从一堆DNA序列中能分离出三种卫星DNA：卫星Ⅰ、Ⅱ和Ⅲ。每一类卫星DNA包括若干不同的串联重复DNA序列家族（卫星亚家族），其中某些是共存于不同类型之中。卫星Ⅱ和Ⅲ大都含有简单的序列重复但也有较高级结构的证据。

αDNA和着丝粒异染色质

其他类型卫星DNA序列不易被密度梯度离心所分解。它们是由限制性内切核酸酶消化基因组DNA后首先被鉴定的，该酶在基本重复单位内具有典型的单个识别位点，除基本重复单位大小（单体）外，由于在某些重复（序列）偶尔随意丢失限制位点，这样的酶会产生以单位长度为特征模式的多聚体（Singer，1982）。α卫星DNA（或

alphoid DNA）由一 171bp 重复单位的串联组成并构成着丝粒异染色质的主体。在 αDNA 家族的每个成员之间的高度序列歧化意味着人类每个染色体存在着特异亚家族（Choo *et al.*,1991）。

卫星 DNA 的确切功能尚不知晓（Csink and Henikoff，1998；Henikoff *et al.*，2001）。人类染色体的着丝粒 DNA 大多由各种卫星 DNA 家族组成（图 9.16）。其中，只有 α 卫星已知存在于所有染色体上，同时其重复单位通常含有一特异着丝粒蛋白，CENP-B 的结合位点。克隆的 α 卫星序列已表明在人类细胞中催生新的着丝粒，表明 α 卫星在着丝粒功能上起着重要作用（Grimes and Cook，1998）。



图 9.16　着丝粒卫星 DNA 组成

不同类型的卫星 DNA 在 9 和 21 号染色体（5 个常染色体端着丝粒之一）上的定位示意，这个插图引自 Tyler-Smith 和 Willard(1993). Curr. Opin. Genet. Dev. 3，390～397，得到 Elsevier 的允许。

9.4.2　小卫星 DNA 是由中等大小排列的串联重复组成，通常位于或靠近端粒

小卫星 DNA（minisatellite DNA）由中等大小排列的串联重复 DNA 序列集合组成，这些序列分散于核基因组的相当的部分（表 9.14）。像卫星 DNA 序列那样，正常情况下它们不转录（例外情况见下文）。

高可变小卫星 DNA（hypervariable minisatellite DNA）序列为高度多态性，并由 1000 以上（从 0.1～20kb 长）排列组成短串联重复（Jeffreys，1987），不同的高可变排列的重复单位，其大小相当不同，但共享一常见的核心序列，GGGCAGGAXG（X＝任一核苷酸），其大小和 G 含量与 Chi 序列（*E. coli* 中一般重组的一个信号）是相似的。而许多这样的排列可见于靠近端粒处，某些高可变小卫星 DNA 序列还发生在其他的染色体部位。虽然大量的高可变小卫星 DNA 序列不转录，但已知在某些少有的情况下是表达的（例如，*MUC1* 基因座，Swallow *et al.*,1987）。

高可变小卫星 DNA 的重要意义仍不清楚，虽然曾报道它是人类细胞中同源重组的“热点”（Wahls *et al.*,1990）。然而，已发现它有许多应用。虽然在亚端粒区的这一优先位置曾限制其在全基因组连锁研究中的应用，但各种单个位点已被鉴定并用作遗传标记。主要用于 DNA 指纹，其中单一 DNA 探针含有常见的核心序列，能与在所有染色体上的多重小卫星 DNA 位点同时杂交，产生复杂的个体特异的杂交模式（节 18.7.1）。

小卫星 DNA 序列的另一主要家族发现于染色体末端的端粒。人类染色体**端粒 DNA**（telomeric DNA）的主要组成是 3～20kb 的串联六核苷酸重复单位，特别是 TTAGGG 重复单位，这是由特异的酶（端粒酶）加上的。其作为一缓冲器通过提供染色体线性 DNA 末端的一种复制机制以保护染色体末端免于降解和丢失。这些简单的重

复直接担负着端粒的功能（图 2.6；节 2.2.5）。

9.4.3 微卫星 DNA 由简单串联重复的短序列组成并分散于全基因组

微卫星 DNA（Microsatellite DNA），也称**简单序列重复**（simple sequence repeat，SSR），是一个简单序列（通常少于 10bp）串联重复的小排列。它们分散于全基因组，占 60Mb 以上（2%的基因组），并认为大多由复制滑动而产生（图 11.5）。双核苷酸重复排列是最常见的类型，约占基因组的 0.5%。CA/TG 重复是很常见的（1/36kb），并常有高度多态性（图 7.7，7.8）。AT/TA（1/50kb）和 AG/CT（1/125kb）重复也十分常见，但 CG/GC 重复则很少（1/10Mb），因为 CpG 双核苷酸易于甲基化并继之去氨基（节 9.1.3）。

在单核苷酸重复之中，A 和 T 的重复很常见（图 9.19 的基因内的例子）；G 和 C 的重复则很少见。三核苷酸和四核苷酸串联重复的单个类型较少，但常有很高的多态性并伴随研究已发现了高度多态的标记。见人类基因组测序协作组（2001）表 14 和 15 的进一步的信息。

微卫星 DNA 的重要意义仍不知道。选择性嘌呤-嘧啶重复，诸如双核苷酸对 CA/TG 的串联重复是能够在体外采用的一种选择性的 DNA 构象，即 Z-DNA，但无证据表明它们在细胞内也是这样。虽然微卫星 DNA 一般在基因间 DNA 中或在基因的内含子中被鉴定，少数例子表明在基因的编码序列中也存在并因为它们易于复制滑动而常为突变热点（图 11.14 的某些例子）和某些有限情况下的不稳定扩张（节 11.5.2 和 16.6.4）。

9.5 散在重复非编码 DNA

9.5.1 转座子衍生重复形成>40%的人类基因组并大多由 RNA 介导产生

人类基因组中几乎所有的散在重复非编码 DNA 都是来自**可转座元件**［transposable element，也称为**转座子**（transposon）］，可移动的 DNA 序列能迁移到基因组的不同区域（Smit，1996；Park and Kazazian，2000）。已识别基因组的约 45%属于这一类（国际人类基因组测序协作组，2001；Li *et al*.，2001），但许多仍为‘单一’DNA，也必定是衍生自古代的转座子拷贝，由于它们分支太远以致这样难以识别。过去，常草率地将它们作为垃圾 DNA，现有不断增加的证据表明这样的转座子对哺乳动物细胞可能是有价值的（Dennis，2002）。

在人类和其他哺乳动物有 4 大类转座子，但只有极少数是活性转座的。按以上转座的方法可分为两组：

- **反转座子**（retrotransposon，也缩写为 retroposon）。这里拷贝的机制运用反转录酶形成 RNA 转录物的 cDNA 拷贝，类似产生加工假基因和反基因的方式（节 9.3.6），**复制的（或拷贝）转座子**［replicative（or copy）transposition］确认一个拷贝是由它随处迁移和插入到基因组后存在的序列所形成的。三类哺乳类的转座子属于这一组：长散在核元件（LINES）；短散在核元件（SINES）和含有长末端重复的反病毒样元件；
- **DNA 转座子**（DNA transposon）。由保守的转座子迁移的第 4 类转座子的成员。此

序列没有拷贝，却是运动的，而后再插入基因组的另一处（一种"切割和粘贴"的机制）。

依其能否独自转座，可分为自主的或非自主的（autonomous or nonautonomous）可转座元件（图 9.17）。四类可转座元件中，LINES 和 SINES 占主要，并将在节 9.5.2 和 9.5.3 分别作更详细的描述。另两类在此予以简要说明。

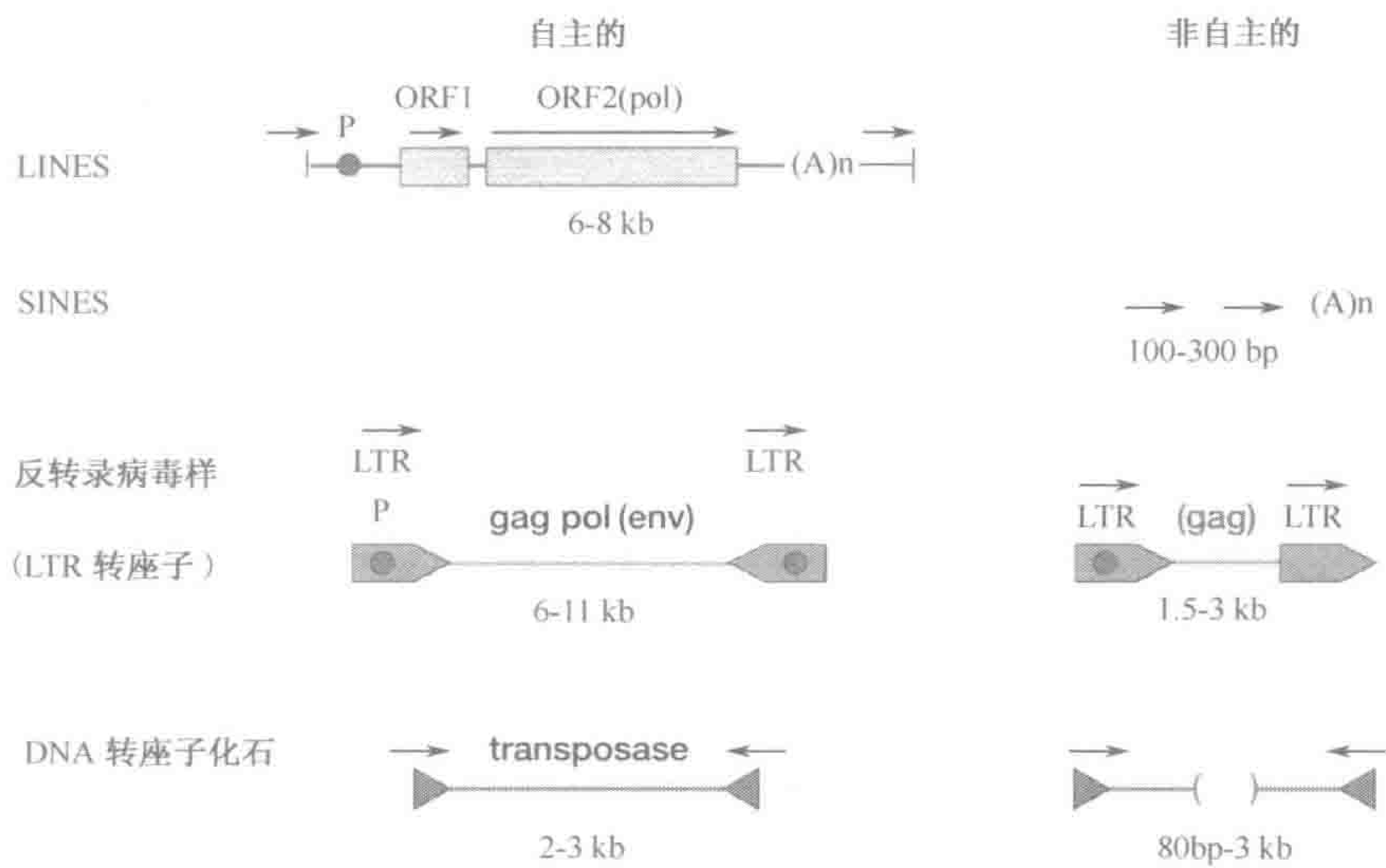


图 9.17　哺乳动物转座子家族

上面任何家族中只有一小部分成员可以转座；许多都由于获得性失活突变而丧失这样的功能，许多是短的截断拷贝。见图 9.18 和 9.19 的某些人类可转座元件的典型结构。经 Nature Publishing Group 的允许，修改自国际人类基因组测序协作组（2001）。Nature 409，860～921。

人类 LTR 转座子

LTR **转座子**（LTR transposon）包括自主的和非自主的反病毒样元件，是由含有必需的转录调节成分的**长末端（直接的）重复**（LTR）构成。已知自主的成员为**同源性反病毒序列**（或 ERV），它们含有 *gag* 和 *pol* 基因，编码蛋白酶、反转录酶、RNAse H 和整合酶。人类 ERV（HERV）有三个主要类型，约占人类基因组的 4.6%（表 9.15）。在最近几百万年中很多是有缺陷并极少转座的，很小的 HERV-K 组表明完整的反病毒基因的保守性（Lower *et al.*,1996），而 HERV-K10 亚家族的某些成员在进化过程的较近年代进行转座。非自主的转座元件缺乏 *pol* 基因而且也常缺乏 *gag* 基因（由于在侧 LTR 之间的同源重组，此内部的序列已经丢失）。MaLR 家族的这些元件几乎占基因组的 4%左右（表 9.15）。

人类 DNA 转座子化石

DNA **转座子**（DNA transposon）具有末端反向重复并编码一调节转座转座酶。它们约占人类基因组的 3%，并能分为不同的类型，每类又可再分为具有独立起源的许多家族（Smit，1996 和 RepBase 重复序列数据库 http://www.girinst.org/）。有两个主

要的人类家族，MER1 和 MER2 加上各种不常见的家族（表 9.15）。所有这些固有的人类 DNA 转座子序列不再有活性，以致成为转座子化石。DNA 转座子在一个物种内趋于短的生命周期，不像其他的某些可转座元件，诸如 LINES（节 9.5.2）。然而，十分少的人类功能基因像是起源于 DNA 转座子，显著地有编码 RAG1 和 RAG2 重组酶的基因和主要的着丝粒结合蛋白 CENPB(Jura and Kapitonov，1999；Smit，1999；国际人类基因组测序协作组，2001)。

表 9.15　人类基因组（除 Y 染色体）中散在重复 DNA 的主要类型和家族

类型	家族	拷贝数	占基因组的百分含量
SINE	Alu 家族	～1 200 000	10.7
	MIR	～450 000	2.5
	MIR3	～85 000	0.4
LINE	LINE-1 家族	～600 000	17.3
	LINE-2 家族	～370 000	3.3
	LINE-3 家族	～44 000	0.3
LTR elements	ERV 家族	～240 000	4.7
	MaLR	～285 000	3.8
DNA transposon	MER-1(Charlie)	～213 000	1.4
	MER-2(Tigger)	～68 000	1.0
	其他	～60 000	0.4

数据来自国际人类基因组测序协作组，2001；小鼠基因组测序协作组，2002。

9.5.2　某些人类 LINE1 元件活跃地转座并能使 SINE 加工假基因和反基因

长散在核元件（Long interspersed nuclear element，LINE）曾是很成功的可转座元件并且具有很长的进化历史。作为自主的可转座元件，它们能编码必需的产物以确保反转座，包括主要的反转录酶。人类 LINES 由三个不同的相关家族组成：LINE1(L1)，LINE2 和 LINE3，集合组成大约 20％的人类基因组（表 9.15）。它们主要位于常染色质区并优先地位于中期染色体的富含 AT 的深染 G 带（Geimsa 阳性）（Korenberg and Rykowski，1988）。在这三个人类家族中，LINE1（或 L1）是仅有的仍然活跃的转座家族，并且它是主要的，构成大约基因组的 17％。它是最重要的人类可转座元件，并也可见于包括小鼠的其他哺乳类（Ostertag and Kazazian，2001）。

LINE1(L1) 元件全长约 6.1kb，编码两个蛋白质：RNA 结合蛋白质和一个具有内切核酸酶和反转录酶活性的蛋白质（图 9.18A）。一个例外的内部启动子位于 5′UTR 内，以致全长转录物的拷贝携有它们自己的启动子，能用于随后在基因组的允许区域内进行整合。在翻译后，LINE1 RNA 与其自己编码的蛋白质组装并移向细胞核。内切核酸酶切割一 DNA 双链，在一条链上留下游离 3′羟基，作为 LINE RNA 的 3′端反转录的引物。内切核酸酶优先切割的位点是 TTTT↓A；因此优先整合到富含 AT 区。富含 AT DNA 在基因中是贫乏的，以致它们趋于整合到富含 AT DNA，这意味着 LINES 给予较低的突变负担，使其宿主易于适应它们。

在整合过程中，反转录通常不能在5′端进行，结果产生截短的无功能插入。因此，大多LINE衍生的重复序列都是短的，所有LINE1拷贝平均大小900bp，而且仅约1/100拷贝是全长的。LINE1装置负责基因组中大多数的反转录，使非自主的SINE反转座并创造加工假基因和反基因（Esnault *et al.*,2000；节9.3.6）。在6000左右的全长LINE1序列中约有60～100仍能转座，并由于随着插入一个重要保守的序列后破坏基因功能，偶尔也引起疾病（节11.5.6）。

9.5.3 在人类基因组中每3kb发生不止一次Alu重复并易受正性选择支配

短散在核元件（short interspersed nuclear element，SINE）大约长100～400bp，并已很成功地在哺乳动物基因组中克隆化，产生许多高拷贝数家族。某些人类SINES是灵长类特异的如Alu家族；其他的不限于灵长类。还发现于有袋目和单孔目，并称为MIR(**m**ammalian-wide **i**nterspersed **r**epeat）家族（表9.15）。SINES不编码任何蛋白质也不是自主性的。LINES和SINES共享它们的3′端的序列，同时SINES已表明通过相邻的伴侣LINES而移动（Kajikawa and Okada，2002)。通过寄生于LINE元件转座装置，SINES能维持很高的拷贝数。

哺乳动物SINES起源自tRNA拷贝（在许多情况下)，或源自SRP（7SL）RNA，正如同Alu重复（Ullu和Tschudi，1984）和小鼠B1重复（节12.4.1）的情况。编码tRNA和SRP RNA的基因由RNA聚合酶Ⅲ转录而它们有其内部启动子是罕见的（图10.4)。然而，由Alu重复携带的内部聚合酶Ⅲ启动子是不能满足体内活性转录的，而适当的侧序列是其激活需要的。因此，随之整合一新的转座Alu拷贝将会失活，除非它偶然地整合于一个能使启动子激活的区域。

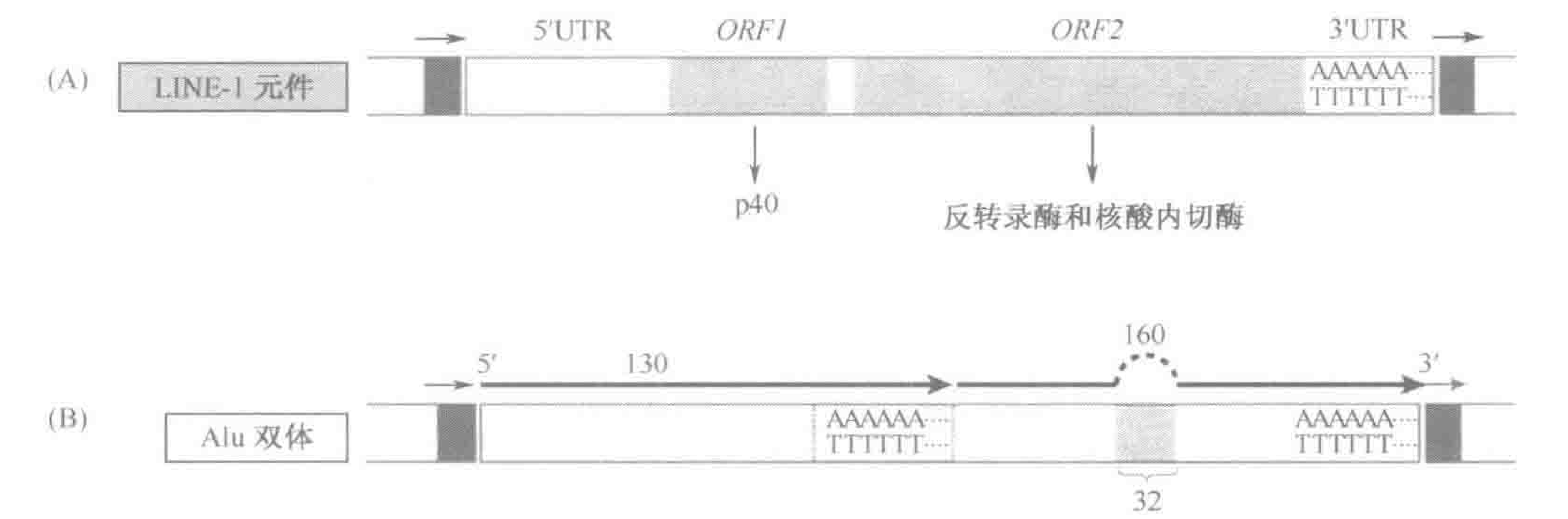


图 9.18　人类LINE1和Alu重复元件

(A) LINE1 (L1) **元件**有两个可读框：1kb的ORF1编码一RNA结合蛋白质，4kb的ORF2编码一具有内切核酸酶和反转录酶活性的蛋白质。一内部启动子位于在ORF1前的非翻译DNA（通常称为5′-UTR）而在另一端有一 $(A)_n/(T)_n$序列。常描述为3′poly(A) 尾。LINE1内切核酸酶切割（↓）DNA双链的一个链，最好在序列TTTT↓A内，同时反转录酶用于主要的cDNA合成时释放3′OH末端。通过一小靶位点在旁侧插入位点复制7～20bp。(B) Alu **重复**。这一致标准的Alu双体表明具有两个相似的以 $(A)_n/(T)_n$样序列为末端的重复。由于在较大的重复序列内插入一32bp元件，故他们的大小不同。Alu单体也存在于人类基因组。就如存在各种截短的单体和双体拷贝。

Alu 重复是人类基因组中最丰富的序列，平均每 3kb 发生不止一次重复（国际人类基因组测序协作组，2001；Li *et al.*,2001）。在不同的进化年代有一系列的 Alu 亚家族，自从人类约在 5 百万年前从非洲类人猿分支以来，只有约 5000 拷贝整合到基因组（Batzer and Deininger，2000）。Alu 重复全长大约 280bp 并由两个串联重复组成，每个约长 120bp，其后为一短序列，它的一链富有 A 残基，而互补链则富有 T 残基。然而，在串联重复之间存在不对称；一重复序列含有一内在的 32bp 序列，而另一重复则缺此序列（图 9.18B）。只含有两个串联重复之一的单体以及含有各种截短副本的双体和单体也是常见的，使之在全基因组平均为 230bp。

Alu 重复有相对高的 GC 含量，虽然主要分散于基因组的全部常染色质区，且主要位于富含 GC 和富含基因的 R 染色体带，而显著对比的 LINE 主要位于富含 AT 的 DNA 区（Korenberg and Rykowski，1988）。然而，当它们位于基因内时，像 LINE1 元件那样，则限于内含子和非翻译区（图 9.19）。尽管趋势是位于富含 GC 的 DNA 区，但新转座的 Alu 重复显示在类似于 LINE 的重复富含 AT DNA 的优势，但是渐进较老的 Alu 显示一种渐进较强的趋向于富含 GC DNA（国际基因组测序协作组，2001）。

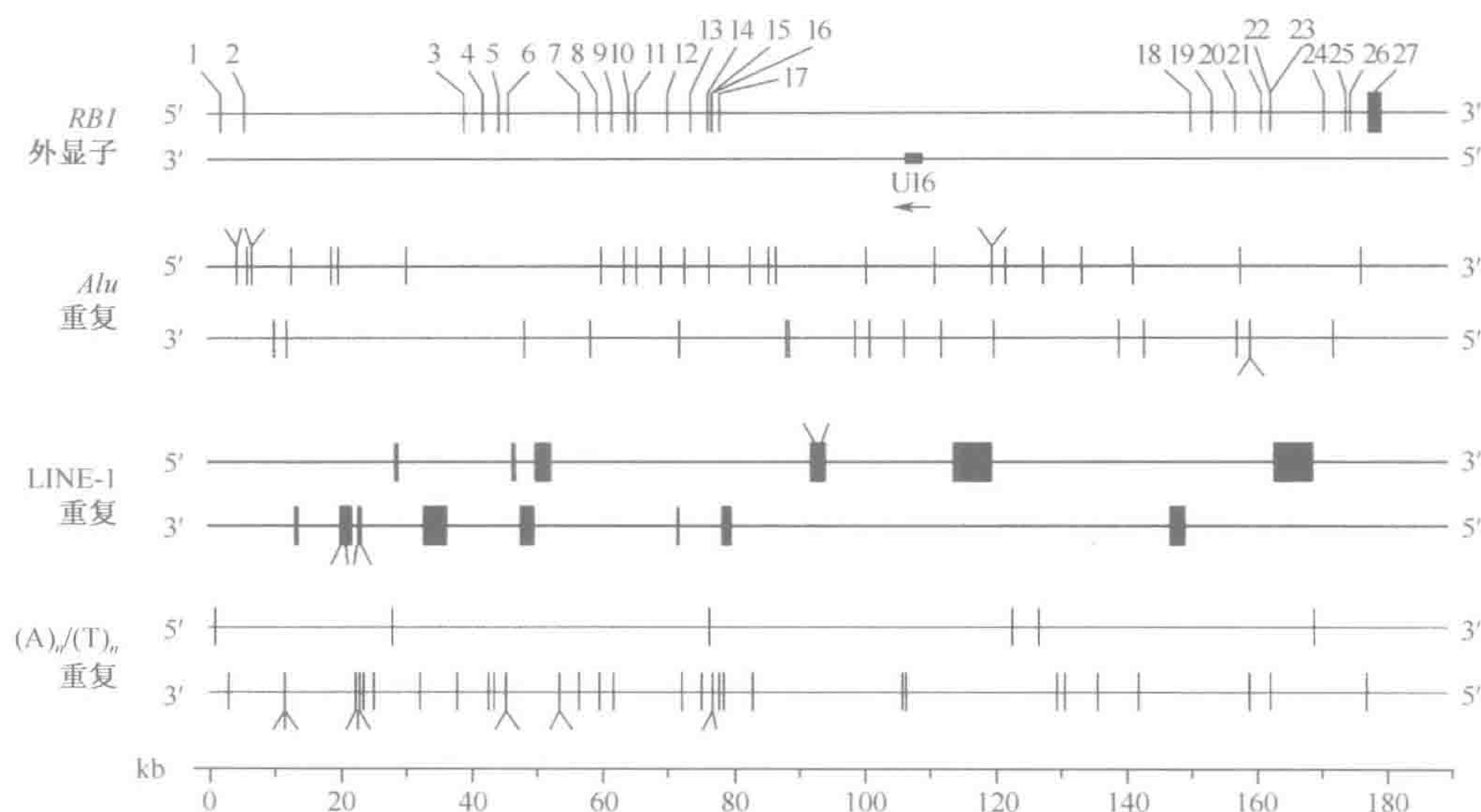


图 9.19 Alu，LINE1 和 $(A)_n/(T)_n$ 重复在人视网膜母细胞瘤易感基因 *RB1* 内的位置

72kb 的内含子 17 含有一 G 蛋白偶联受体基因，U16，它以 *RB1* 基因的反方向主动转录，每对的上线（5′→3′）表示重复元件在 *RB1* 基因有义链的方向。下线（3′→5′）表示它们在反义链的方向。有 46 Alu 重复和 17 LINE1 元件（有些紧密簇集用分支线表示），全都位于内含子。只有两个 LINE1 元件全长达 6.1kb。$(A)_n/(T)_n$ 的序列，（n=12 或更多）表明只是那些位于散在重复之外的。未发现 n=12 或更多的 $(C)_n/(G)_n$ 的例子。重画自 Toguchida *et al.*，(1993). Genomics 17，535～543，经 Elsevier 允许。

Alu 整体分布趋向富含 GC（并因此也趋向富含基因）区必定是由于强烈的选择压力所致。这说明 Alu 重复并不是基因组的寄生物，而正在成为细胞所含有的一个有用的内容物。某些 Alu 序列一直是活跃转录的并可能已补充为一种有用的功能。*BCYRN1* 基因编码一天然的神经小细胞质 RNA，BC200，由一 Alu 单体产生并是少数的 Alu 序列之一，在正常环境下，该序列具有转录活性（Martignetti and Brosius，

1993)。在许多物种中，SINE 在压力条件下转录，而由此产生的 RNA 结合一特异的蛋白质激酶（PKR），阻遏其能力以抑制蛋白质翻译。因此，SINE RNA 在压力下促进蛋白质翻译。所以，很可能 SINES 的一般功能（Schmid，1998）就在于调节蛋白质翻译（SINE RNA 以上千的元件大量很快地转录并能在没有蛋白质翻译下发挥作用）。

（孙开来 译）

进一步阅读

Human Genome *Nature* Issue (15 February 2001). *Nature* **409**, 813–958 (papers are available electronically via the Nature Genome Gateway at http://www.nature.com/genomics/human/)

Human Genome *Science* Issue (16 February 2001). *Science* **291**, 1177–1351 (papers are available electronically at http://www.sciencemag.org/content/vol291/issue5507/index.shtml)

InterPro proteome analysis database at http://www.ebi.ac.uk/proteome/

MITOMAP human mitochondrial genome database at http://www.mitomap.org

Mouse Genome *Nature* Issue (5 December 2002). *Nature* **420**, 447–590 (papers are available electronically via the Nature Genome Gateway at http://www.nature.com/nature/mousegenome/index.html)

NCBI guide to on-line information resources on the human genome at http://www.ncbi.nlm.nih.gov/genome/guide/human/

Noncoding RNAs Database at http://biobases.ibch.poznan.pl/ncRNA/

Pfam protein domain family database at http://www.sanger.ac.uk/Software/Pfam/

Repeat Sequence Database at http://www.girinst.org

参考文献

Adachi N, Lieber MR (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807–809.

Ambros V (2001) microRNAs: tiny regulators with great potential. *Cell* **107**, 823–826.

Anderson S, Bankier AT, Barrell BG *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.

Bailey JA, Gu Z, Clark RA *et al.* (2002) Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.

Batzer MA, Deininger PL (2002) Alu repeats and human genetic diversity. *Nature Rev. Genet.* **3**, 370–378.

***C. elegans* sequencing consortium** (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012–2018.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. *Nature Genet.* **31**, 415–418.

Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P (1991) A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucl. Acids Res.* **19**, 1179–1182.

Claverie JM (2001) Gene number. What if there are only 30,000 human genes? *Science* **291**, 1255–1257.

Clayton DA (1992) Transcription and replication of animal mitochondrial DNAs. *Int. Rev. Cytol.* **141**, 217–232.

Collins JE, Goward ME, Cole CG *et al.* (2003) Reevaluating human gene annotation: a second-generation analysis of chromosome 22. *Genome Res.* **13**, 27–36.

Craig JM, Bickmore WA (1994) The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7**, 376–381.

Csink AK, Henikoff S (1998) Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.* **14**, 200–204.

Dennis C (2002) A forage in the junkyard. *Nature* **420**, 458–459.

Dermitzakis ET, Reymond A, Lyle R *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420**, 578–582.

Eddy SR (2001) Noncoding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929.

Eichler EE (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.* **17**, 661–669.

Esnault C, Maestre J, Heidmann T (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24**, 363–367.

European Polycystic Kidney Disease Consortium (1994) The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome **16**. *Cell* **77**, 881–894.

FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573.

Filipowicz W (2000) Imprinted expression of small nucleolar RNAs in brain: time for RNomics. *Proc. Natl Acad. Sci. USA*, **97**, 14035–14037.

Frugier T, Nicole S, Cifuentes-Diaz C, Melki J (2002) The molecular bases of spinal muscular atrophy. *Curr. Opin. Genet. Dev.* **12**, 294–298.

Geraghty DE, Koller BH, Pei J, Hansen JA (1992) Examination of four HLA class I pseudogenes. Common events in the evolution of HLA genes and pseudogenes. *J. Immunol.* **149**, 1947–1956.

Gottesman S (2002) Stealth regulation: biological circuits with small RNA switches. *Genes. Dev.* **16**, 2829–2842.

Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. *Genome Res.* **11**, 685–702.

Gray TA, Saitoh S, Nicholls RD (1999) An imprinted mammalian bicistronic transcript encodes two independent proteins. *Proc. Natl Acad. Sci. USA* **96**, 5616–5621.

Grimes B, Cooke H (1998) Engineering mammalian chromosomes. *Hum. Mol. Genet.* **7**, 1635–1640.

Harrison PM, Hegyi H, Balasubramanian S *et al.* (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**, 272–280.

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102.

Huttenhofer A, Brosius J, Bachellerie JP (2002) RNomics: identification and function of small, non messenger RNAs. *Curr. Opin. Chem. Biol.* **6**, 835–843.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Jeffreys AJ (1987) Highly variable minisatellites and DNA fingerprints. *Biochem. Soc. Trans.* **15**, 309–317.

Jurka J, Kapitonov VV (1999) Sectorial mutagenesis by transposable elements. *Genetica* **107**, 239–248.

Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111**, 433–444.

Kapranov P, Cawley SE, Drenkow J *et al.* (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919.

Korenberg JR, Rykowski MC (1988) Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858.

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T (2002) Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* **12**, 735–739.

Lanz RB, McKenna NJ, Onate SA *et al.* (1999) A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* **97**, 17–27.

Lehner B, Williams G, Campbell RD, Sanderson CM (2002) Antisense transcripts in the human genome. *Trends Genet.* **18**, 63–65.

Li W-H, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature* **409**, 847–849.

Locascio A, Vega S, de Frutos CA, Manzanares M, Nieto MA (2002) Biological potential of a functional human SNAIL retrogene. *J. Biol. Chem.* **277**, 38803–38809.

Lower R, Lower J, Kurth R (1996) The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc. Natl Acad. Sci. USA* **93**, 5177–5184.

Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243.

Martignetti JA, Brosius J (1993) BC200 RNA: a neural RNA polymerase III product encoded by a monomeric Alu element. *Proc. Natl Acad. Sci. USA* **90**, 11563–11567.

Mefford HC, Trask BJ (2002) The complex structure and dynamic evolution of human subtelomeres. *Nature Rev. Genet.* **3**, 91–102.

Mourelatos Z, Dostie J, Paushkin S *et al.* (2002) miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.* **16**, 720–728.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Nei M, Rogozin IB, Piontkivska H (2000) Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl Acad. Sci. USA* **97**, 10866–10871.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA (2000) The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**, 920–930.

Ostertag EM, Kazazian HH Jr (2001) Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538.

Prak EL, Kazazian HH Jr (2000) Mobile elements and the human genome. *Nature Rev. Genet.* **1**, 134–144.

Regnier V, Meddeb M, Lecointre G *et al.* (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6**, 9–16.

Schmid CW (1998) Does SINE evolution preclude Alu function? *Nucl. Acids Res.* **26**, 4541–4550.

Schmid SR, Linder P (1992) D-E-A-D protein family of putative RNA helicases. *Mol. Microbiol.* **6**, 283–291.

Singer MF (1982) Highly repeated sequences in mammalian genomes. *Int. Rev. Cytol.* **76**, 67–112.

Smit AF (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**, 743–748.

Smit, AF (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663.

Smith CM, Steitz JA (1997) Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell* **89**, 669–672.

Smith TF, Gaitatzes C, Saxena K, Neer EJ (1999) The WD-repeat: a common architecture for diverse functions. *Trends Biochem. Sci.* **24**, 181–184.

Storz G (2002) An expanding universe of noncoding RNAs. *Science* **296**, 1260–1263.

Swallow DM, Gendler S, Griffiths B, Corney G, Taylor-Papadimitriou J, Bramwell ME (1987) The human tumour-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. *Nature* **328**, 82–84.

Tennyson CN, Klamut HJ, Worton RG (1995) The human dystrophin gene requires 16 hours to be transcribed and is co-transcriptionally spliced. *Nature Genet.* **9**, 184–190.

Toguchida J, McGee TL, Paterson JC, Eagle JR, Tucker S, Yandell DW, Dryja TP (1993) Complete genomic sequence of the human retinoblastoma susceptibility gene. *Genomics* **17**, 535–543.

Tycowski KT, Shu M-D, Steitz JA (1993) A small nucleolar RNA is processed from an intron of the human gene encoding ribosomal protein S3. *Genes Dev.* **7**, 1176–1190.

Tyler-Smith C, Willard HF (1993) Mammalian chromosome structure. *Curr. Opin. Genet. Dev.* **3**, 390–397.

Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. *Nature* **312**, 171–172.

Valadkhan S, Manley JL (2001) Splicing-related catalysis by protein-free snRNAs. *Nature* **413**, 701–707.

Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

Wahls WP, Wallace LJ, Moore PD (1990) Hypervariable minisatellite DNA is a hotspot for homologous recombination in human cells. *Cell* **60**, 95–103.

Yang Z, Zhu Q, Luo K, Zhou Q (2001) The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**, 317–322.

Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, Trask BJ (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* **11**, 535–546.

Zhang MQ (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**, 919–932.

Zhang MQ (2002) Computational prediction of eukaryotic protein-coding genes. *Nature Rev. Genet.* **3**, 698–709.

Zhang Z, Harrison P, Gerstein M (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12**, 1466–1482.

第 10 章　人类基因表达

本 章 内 容

10.1　人类细胞中基因表达概述

用于调节人类基因表达的控制机制可能比低等真核生物中更为复杂，但许多相同的基本规则仍然适用。哺乳动物是尤其复杂得多细胞有机体，因此也许不足为奇的是部分控制哺乳动物基因表达的机制并未在细菌或其他真核生物中使用。表达具有空间和时间上的局限（框 10.1）。尽管过于简单化，但从三个主要的层次来考虑基因调控的运作却很方便：

- **基因表达的转录调控**。真核细胞中基因表达的初级调控发生于通过基因的核心启动子的转录起始水平以及相关 RNA 聚合酶的聚集和加工水平。基因的表达由转录因子结合至启动子开始。基础的转录水平可通过蛋白因子与旁侧或内含子序列中的其他调控序列相结合来调节；
- **基因表达的转录后调控**。这包括作用于 RNA 加工、mRNA 转运、翻译、mRNA 稳定性、蛋白质加工、蛋白质靶向输送，蛋白质稳定性等水平的机制。在 RNA 加工的水平，不同的机制如 RNA 剪接（一种真正的共转录机制）使一个基因能够产生多种不同的基因产物（**异构体**，isoform）。翻译的调控通常涉及反式蛋白因子对非翻译区调控序列的识别。

▶ **表观遗传机制与基因表达的远程调控**。遗传因子依赖于DNA序列的改变。另外的可遗传（从细胞到子细胞，或者从亲代到子女）、但并不依赖于基因组序列改变的变化被称为表观遗传。DNA甲基化是已知的极少数在哺乳动物细胞中起作用的表观遗传机制之一，在基因表达的调控中扮演重要角色（Bird，2002）。除了作为维持转录抑制的一种普遍方式外，它亦作用于一些基因上、确保两个继承自亲代的等位基因中仅一个正常表达（**单等位性表达**，monoallelic expression）的机制中起关键作用，即不表达的等位基因的核苷酸序列可能与表达者一致。表观遗传机制常导致跨越长距离的染色质构象改变的永存。

框10.1 哺乳动物细胞中基因表达的时空局限性

基因表达的空间局限性

持家基因（housekeeping gene）需要在几乎所有类型的有核细胞中表达，因为它们编码某种在所有细胞中执行一种常规功能，如蛋白质合成、能量产生等所需的关键产物。然而，许多真核基因则呈现局限得多的组织特异性基因表达模式。基因表达的空间限制可发生于不同的层次：

多器官/组织模式。在某些情况下，一个基因可能在不同的器官系统中扮演类似的角色。多种在早期发育中扮演关键角色的基因可参与调节几个不同器官系统的靶基因，例如，sonic hedgehog基因表达于发育期神经系统的各个部分、发育期的四肢以及其他部位。在其他情况下，一个基因可通过使用组织特异性启动子或组织特异性选择性剪接在不同的组织中编码不同的变异体（异构体）。在某些情况下，这些可能具有不同的功能（节10.3.2）；

特定的组织、细胞谱系或细胞类型。一些基因具有适于特定细胞类型或细胞谱系的功能，如同表达于红细胞系统的β珠蛋白基因；

单个细胞。一些特化的基因在属于同一细胞类型的个体细胞中将产生不同的产物。例如，一个人体内不同的B淋巴细胞将表达不同的（细胞特异性）抗体分子，不同的T细胞将产生不同T细胞受体，而单个的嗅觉神经元将产生不同的嗅觉受体。**注意**：由于随机单等位性表达（如在X染色体失活的情况，节10.5.6），在同一组织中相同类型的不同细胞中也可能存在表达的差异（如同X染色体失活，节10.5.6）；

细胞内分布。不同基因的蛋白质将被运输至细胞内（或细胞外）的不同部位。在一些情况下，相同蛋白质的不同异构体可被运送到细胞内的不同部位（节10.3.2）。另外，将一些基因的mRNA转运到细胞内的不同部位将需要基因调控机制（节10.2.6）。

基因表达的时间局限性

细胞周期阶段。除了在有丝分裂期染色体浓缩时普遍的大范围基因沉默以外，一些基因仅表达于细胞周期的特定时间。例如，许多组蛋白基因仅在S(DNA合成）期表达，而各种细胞周期调节因子的表达则被程序化而出现在特定的细胞周期阶段；

发育阶段。在发育的极早阶段不发生转录；相反，细胞将依赖原先合成的RNA。在发育过程稍晚一些，基因可在特定的阶段短暂表达。一些基因家族包含一些在不同发育阶段表达的成员，如珠蛋白基因（图10.22，10.23）；

分化阶段。当细胞分化时，它们的基因组被修饰从而产生改变的基因表达模式。在一些终末分化的细胞中，转录将不发生。导致向有核成年体细胞进展的基因组修饰过去常认为是不可逆的，直到克隆羊Dolly的诞生（节20.2.2）；

可诱导表达。一些基因将响应环境刺激或来自于其他细胞的细胞外信号而激活（节10.2.5）。如果诱导因素被去除，这类基因表达很容易被逆转。

表 10.1 为已知涉及人类基因表达调控的不同类型机制提供了一个概括。

表 10.1　人类细胞中基因表达调节一览

选择性表达机制	例子
转录	
组蛋白修饰，染色质重构	节 10.2.1
组织特异性转录因子与单个基因的顺式作用元件相结合	表 10.3
激素、生长因子或中介物与可诱导转录元件中的反应元件直接结合	cAMP 反应元件，类固醇激素反应元件等(表 10.4；图 10.10)
单个基因不同启动子的使用	肌营养不良基因(图 10.14)；*Dnmt1* 基因(图 10.20)
转录后	
选择性剪接	节 10.3.2；图 10.15 和 10.16
选择性多聚腺苷酸化	节 10.3.2；图 10.15E
组织特异性 RNA 编辑	节 10.3.3；图 10.17
翻译调控机制	节 10.2.6；图 10.13；图 9.6
表观遗传机制/由染色质结构造成的远程调控	
等位基因排斥	产生细胞特异性免疫球蛋白和 T 细胞受体的 B 和 T 淋巴细胞中的 DNA 重排(节 10.6.1)
	表达于女性细胞中的一条 X 染色体上的许多基因被 *XIST* 基因产物失活(节 10.5.6)
	由不明机制造成的随机性等位基因排斥，如 *IL-2*、*IL-4*、*PAX5* 等(节 10.5.3；框 10.4)
	特定基因的印记(节 10.5.4、10.5.5)
由染色质结构造成的远程调控	对于增强子或沉默子的竞争(例如在珠蛋白的表达中；节 10.5.4)
	位置效应(节 10.5.1)
细胞位置依赖性短距离信号	节 10.4.1

10.2　反式作用蛋白因子与 DNA 和 RNA 中的顺式作用调节序列的结合对基因表达的调控

基因表达的调控（无论发生于转录起始、RNA 加工、翻译、RNA 转运等任何水平）在很大程度上均涉及蛋白因子与核苷酸调节序列的结合。后者可以是存在于基因附近乃至于其内部的 DNA 序列，或者是前体 RNA 或 mRNA 水平的转录物序列。由于参与基因表达调节的蛋白因子本身是由远隔的基因编码，它们需要迁移至其作用的位点，因此被称为**反式作用**（*trans*-acting）因子。相反，它们所结合的调节序列通常为**顺式作用**（*cis*-acting），因为它们与被调节的基因或 RNA 转录物位于相同的 DNA 或 RNA 分子上。

通过 DNA-蛋白质结合的调控

在真核细胞中，一个主要的基因表达调控点是在转录起始水平。染色质是一种高度

组织并致密包裹的结构，使 RNA 聚合酶不易与之结合，因此改变染色质的折叠与基本结构需要染色质重构酶，使之成为更为宽松和活跃的结构而使转录能够发生。三种不同类型的 RNA 聚合酶据知可转录不同种类的基因（表 1.4），每种都是由 8～14 个亚基组成的大型酶。

蛋白质（**转录因子**，factor）与基因内或邻近区域的特定 DNA 调节序列结合后，RNA 聚合酶开始转录基因。转录因子可以分为两种：

- 常规转录因子是转录一类特异性 RNA 聚合酶的大多数启动子所必需 [也就是说，常规转录因子多半为一类聚合酶所独有，尽管至少有一种因子，TATA **框结合蛋白**（TATA box-binding protein）为全部三种聚合酶所共有]。对于编码多肽的基因而言，例如一种特定的聚合酶，RNA 聚合酶Ⅱ，与相关的常规转录因子协同作用以产生基础水平的转录。
- 特化转录因子可调控基础表达水平，并包括**组织特异性转录因子**（tissue-specific transcription factor）和特定基因群的转录相关的因子。与特定 DNA 序列结合并刺激转录的蛋白质被称作**激活因子**（activator，或**反式激活因子**，*trans*-activator）。那些具有使转录活性沉默的拮抗效应者被称为**阻遏物**（repressor）。

除了在结合 DNA 之后能够激活或抑制基因表达的转录因子外，种类广泛的调控蛋白将影响基因表达，但并不与 DNA 本身、而是与转录因子自己结合。这类蛋白质被称作**辅激活因子**（co-activator）或**辅阻遏物**（co-repressor）（根据它们所结合的是转录激活因子或阻遏物；Lemon and Tjian，2000）。

通过 RNA-蛋白质结合的调控

除转录因子外，RNA 结合蛋白也被用于调节基因表达。研究得最为清楚的例子涉及与 mRNA 非翻译序列中的调控序列相结合，使基因表达得以在翻译水平被调控。另外，特异的 RNA-蛋白质结合的相互作用被认为在差异性 RNA 加工水平上参与基因表达的调控，如同 SR 和 HnRNP 蛋白与前体 mRNA 结合从而在剪接中调节外显子的选择。后一种机制将在节 10.3 中单独讨论，以例证可用于解码单个基因的表达机制的极大复杂性以及由此产生的大量的异构体的意义。

10.2.1　组蛋白修饰和染色质重构有助于 DNA 结合因子与染色质的结合

非活跃转录的 DNA 被组织成为浓缩的染色质结构，其中核心组蛋白与 DNA 之间具有紧密的联系。核小体的致密包装能够阻止 DNA 相互作用所需的多种不同蛋白质的接近，不仅包括涉及基因表达的蛋白，还包括 DNA 复制、DNA 修复等所需的因子。局部的染色质结构能够通过组蛋白修饰和染色质重构可逆地由浓缩状态转变为一种更易于接近的构象。

组蛋白修饰

组蛋白在核小体中是以如此的方式组织起来，以至于它们的尾部（特别是 N 端）从组蛋白八聚体中伸出来。四个核心组蛋白的暴露的 N 端尾部具有极为高度保守的序列，并在调控染色质结构中执行关键功能。特别是组蛋白 H3 和 H4 已知包含特定可被

修饰的氨基酸（Goll and Bestor，2002 及图 10.1）。在特定染色质区域内组蛋白残基修饰所得到的总体模式被认为形成了一种规定了某种生物学结局的**组蛋白码**（histone code）（Berger，2002；Turner，2002）。

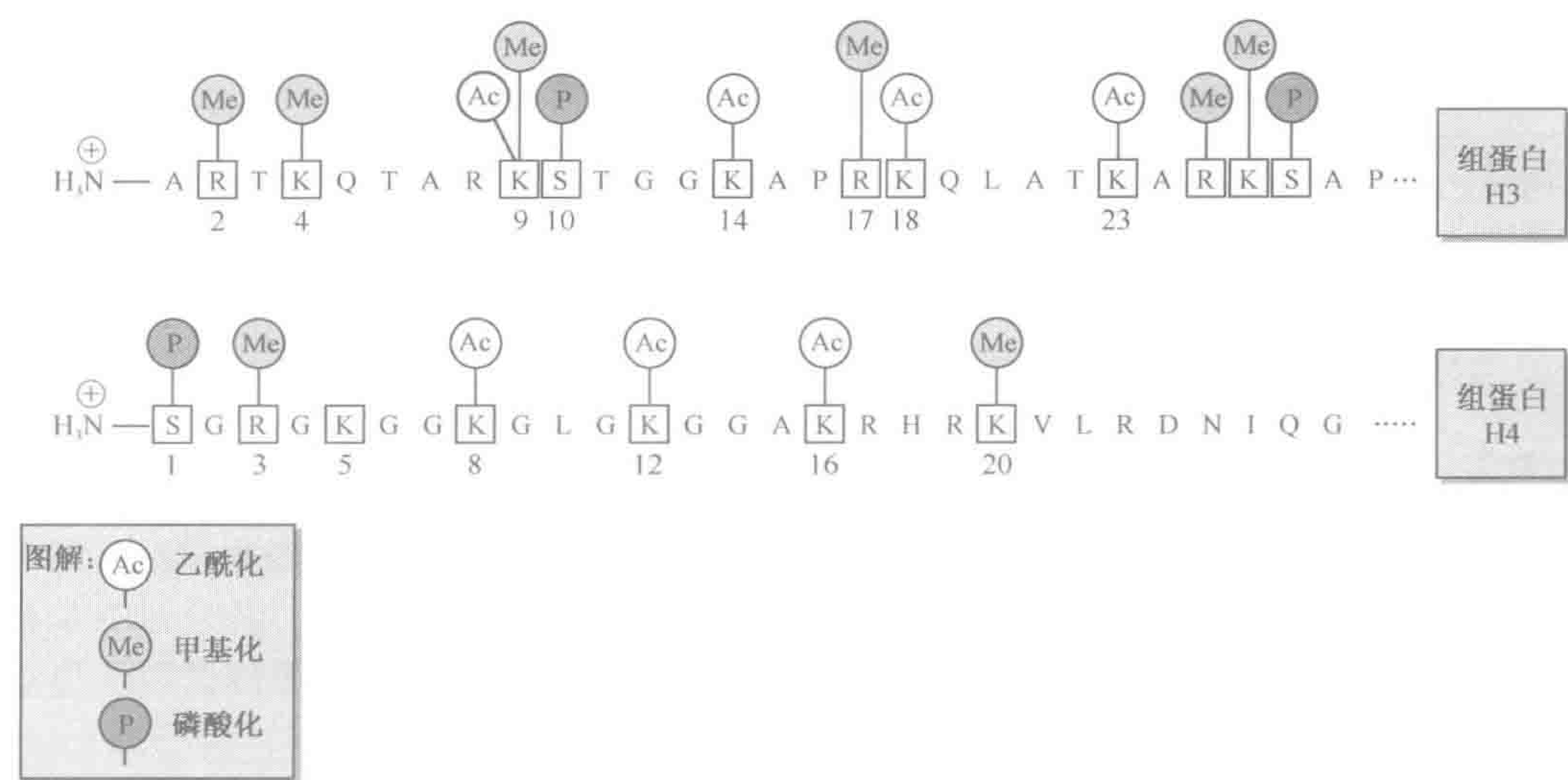


图 10.1　组蛋白 H3 和 H4 的修饰

组蛋白 3 和 4 的 N 端残基（在组蛋白八聚体中时为暴露的尾部的一部分）的修饰如图所示。注意 H3K9（组蛋白 H3 第 9 位置上的赖氨酸残基）既可被乙酰化，又可被甲基化。乙酰化与具有转录活性的染色质相关，但如果染色质的这个区域在 CpG 岛被甲基化，与 CpG 结合的蛋白能够聚集组蛋白脱乙酰酶，后者将造成乙酰基的去除，选择 H3K9 残基将被结合到 CpG 结合蛋白的组蛋白甲基化酶甲基化。这样甲基化的 H3K9 残基就成为诱导染色质浓缩的蛋白质结合的靶点（图 10.21）。

组蛋白乙酰化（histone acetylation）是研究得最为深入的组蛋白修饰（后者还包括甲基化、磷酸化和泛素化）。据知各种组蛋白乙酰基转移酶（histone acetyltransferases，HAT）可催化乙酰基团（CH_3CO^-）加至 30 个赖氨酸残基中多达 13 个的侧链上 ε-NH_3^+ 氨基基团，这些侧链暴露在组蛋白八聚体的 N 端尾部。HAT 作为转录辅活化物发挥作用。在一定程度上，这可能是由于带电荷的赖氨酸侧链被修饰时正电荷的丢失，造成组蛋白同 DNA 亲和力的下降。修饰的组蛋白也是染色质重构机制的作用对象（见下面）。其净效应为 RNA 聚合酶及转录因子更易于接近启动子区域。**组蛋白脱乙酰酶**（histone deacetylase，HDAC）具有去除乙酰基团的相反效应，从而推动转录抑制。HDAC 被认为是响应 DNA 甲基化而聚集的辅抑制物复合体的一部分（节 10.4.3；图 10.21）。

组蛋白甲基化（histone methylation）由不同类型的以组蛋白内特定的精氨酸以及特定的赖氨酸残基为目标的甲基转移酶来实现。组蛋白精氨酸的甲基化据知参与转录激活，而组蛋白赖氨酸甲基化则可能是转录抑制的信号，如可通过相同残基去乙酰化而诱导的 H3K9（组蛋白 3，第 9 位赖氨酸）甲基化（节 10.4.2；图 10.21）。

染色质重构

ATP 依赖性**染色质重构复合体**（chromatin remodeling complexe）通过 ATP 水解

而暂时改变核小体的结构（Narlikar *et al*.,2002）。通过染色质重构，各种蛋白质将更易于与 DNA 接近。即便在重构复合体自 DNA 分离之后，核小体还可以在一段时期内保持 DNA-组蛋白接触被松解的重构状态。在一些情况下，重构还可以涉及核小体位置的改变，导致它们沿 DNA 滑动。作为核小体滑动的结果，原先缠绕组蛋白八聚体的序列变得更容易接近。另外，某些重构复合体能够恢复转录失活的状态。

染色质重构复合体种类繁多，每种均包括多个亚基（通常＞10）。除具有 ATP 酶结构域之外，它们还展现出其他与修饰的组蛋白相互作用的结构域，使组蛋白修饰与染色质重构得以协调进行。它们包括 SW1/SNF 家族的**溴区结构域**（bromodomain），后者据知与组蛋白内乙酰化的赖氨酸残基相互作用，以及与甲基化的赖氨酸相互作用的 Mi-2 家族的**染色质结构域**（chromodomain）（Berger，2002；Narlikar *et al*.,2002）。

10.2.2　由 RNA 聚合酶Ⅰ和Ⅲ进行的转录需要泛在的转录因子

真核细胞中的 RNA 聚合酶Ⅰ和Ⅲ被专用于转录基因从而产生帮助编码多肽基因表达的 RNA 分子（rRNA，tRNA 等）。这些转录的基因属于持家基因，因为 rRNA 和 tRNA 几乎为所有的细胞需要以辅助蛋白质的合成。因此，需要泛在的转录因子协助 RNA 聚合酶Ⅰ和Ⅲ。

由 RNA 聚合酶Ⅰ进行的转录

RNA 聚合酶Ⅰ仅局限于核仁中，专用于 18S、5.8S 和 28S rRNA 基因的转录。后者连续地分布于一条共同的 13 kb **多顺反子转录单位**（polycistronic transcription unit）上（图 10.2）。由 13 kb 的转录单位和邻近的 27 kb 非转录**间隔区**（spacer）组成的复合单元在 5 条人类近端着丝粒染色体每一条短臂上的**核仁组织区**（nucleolar organizer region）将串连重复 30～40 次。所产生的 5 个 rRNA 基因簇，每个长约 1.5 Mb，被称作**核糖体** DNA 或 rDNA。

随着两个转录因子与某个核心的启动子元件在转录起始位点以及一个位于上游超过 100 个核苷酸处的**上游调控元件**（up stream factor）相结合，28S、5.8S 和 18SrRNA 的转录将开始。转录因子之一，UBF（upstream binding factor，**上游结合因子**）为同型二聚体，它的两个完全相同的亚基可能首先结合至核心启动子和上游调控元件，使二者凑到一起，从而使它们能够被第二种因子，SL1（选择性因子 1；在小鼠中据知为 TIF-1B；图 10.3）结合。结合的转录因子随后将聚集 RNA 聚合酶Ⅰ从而形成起始复合体。

由单个 13 kb 转录单位表达形成的初级转录物是一种 45S 的 rRNA 前体，后者将经过多种剪接反应与碱基特异性修饰（由大量不同种类的小核仁 RNA 完成），从而产生成熟的 28S、5.8S 和 18S rRNA（图 10.2）。因此，这些基因与个别转录的绝大多数核基因不同。相反，rDNA 的转录与 mtDNA 相似（节 9.1.2；图 9.2）：二者均产生能够形成功能相关产物的多基因转录物。然而，这种非同寻常的多基因初级转录物的使用在原理上与单一初级翻译产物偶尔被剪接产生两种或更多的功能相关多肽并无不同（见图 1.23 中人类胰岛素的例子）。

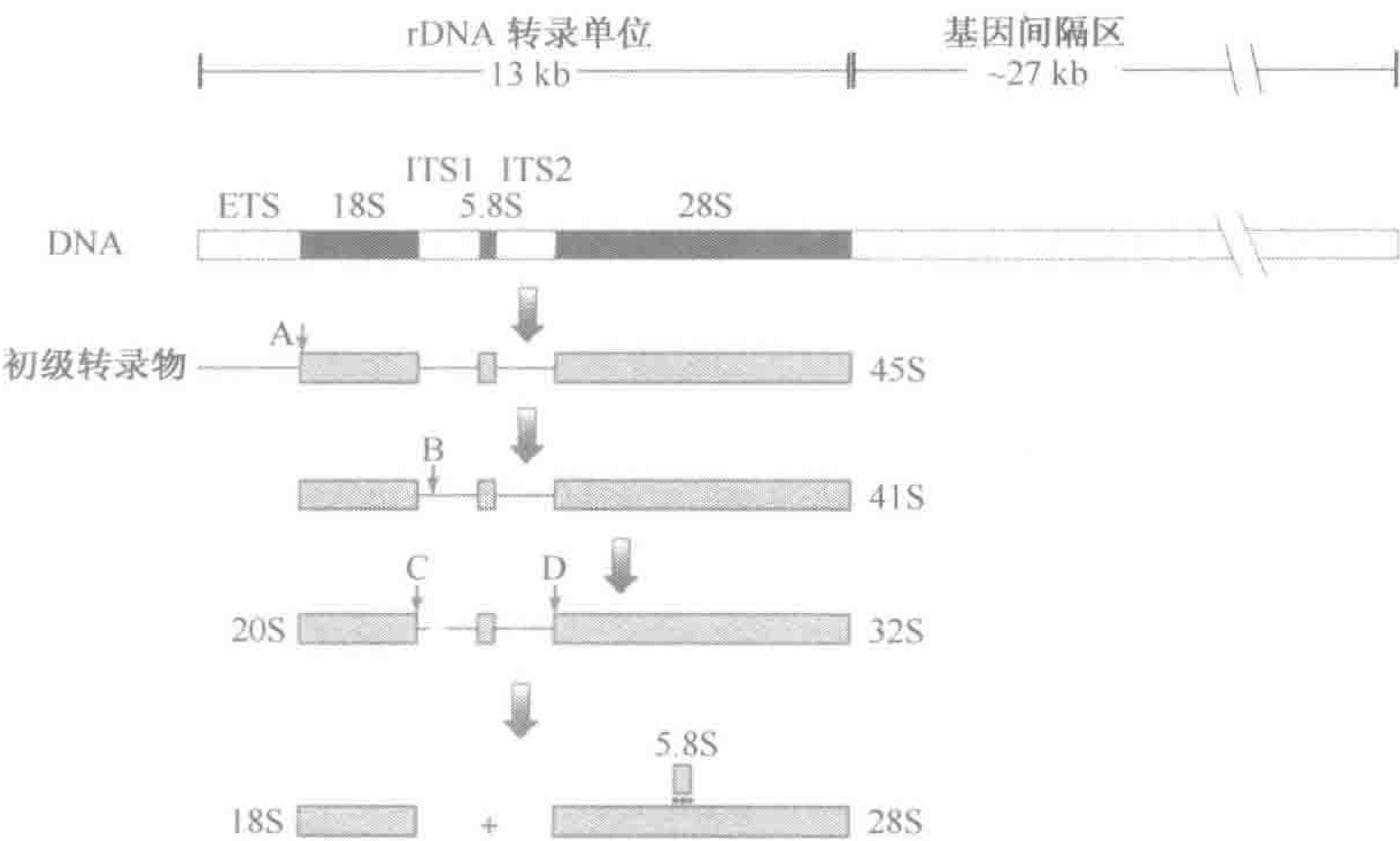


图 10.2 人类主要的 rRNA 种类是由切割一条共同的 13 kb 转录单位而合成的，后者为一条 40 kb 的串联重复单元的部分

字母 A～D 所示的短箭头表示 RNA 前体被内切核酸酶切割的位置。41S 前体在 B 位点切开后将产生两种产物：20S＋32S。随着 32S 前体在 D 位点的切割和小的 5.8S rRNA 的切除，在 5.8S rRNA 和 28S rRNA 一个互补的中心片段之间将形成氢键。由外部和内部转录间隔单元（ETS、ITS1 和 ITS2）所产生的约 6 kb 的 RNA 序列在核内降解。S 为沉降系数，为分子大小的度量。

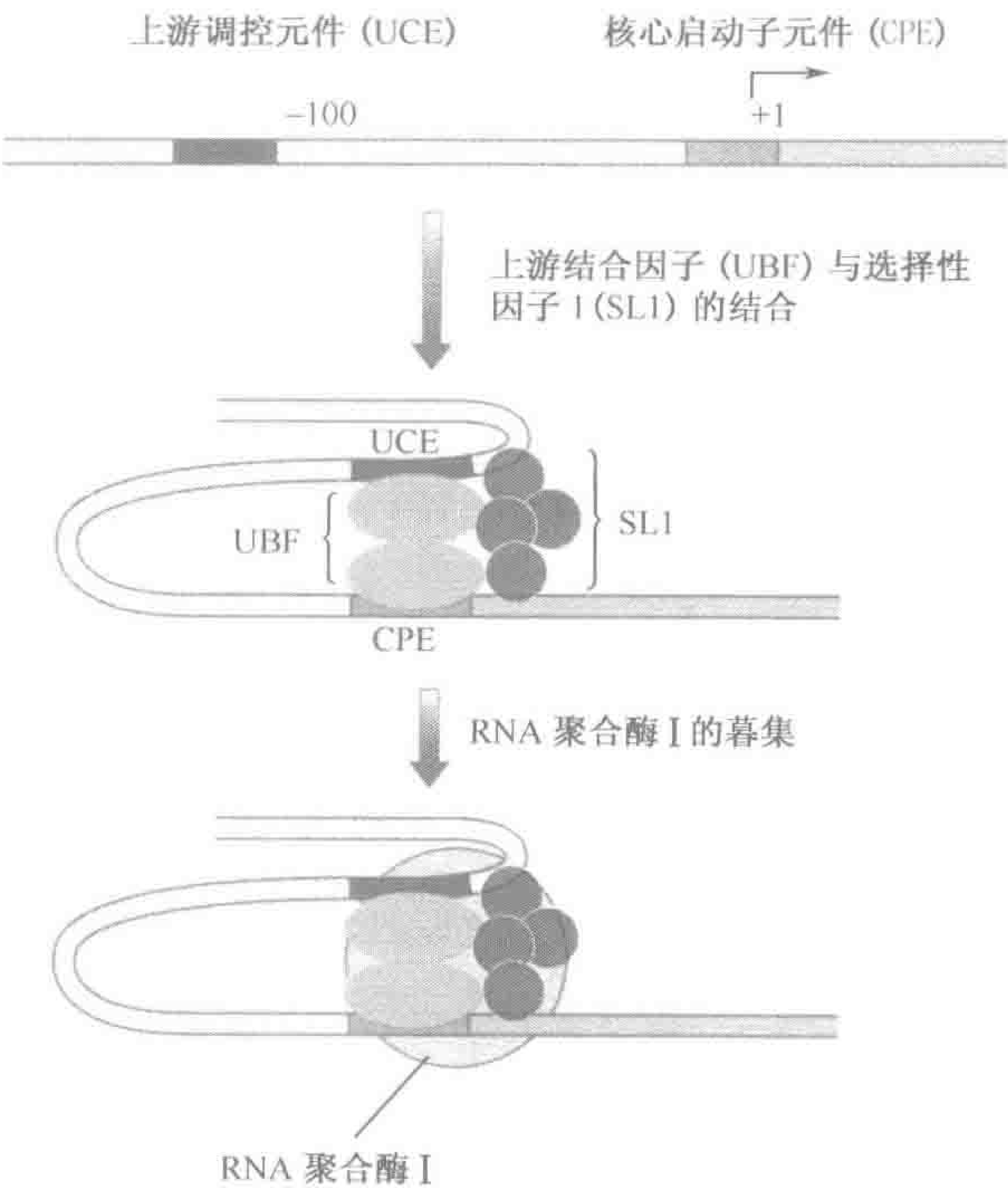


图 10.3 由 RNA 聚合酶 I 起始的转录

一种可能的模型推测上游结合因子的两个相同亚基最初与上游的调控元件和核心启动子元件相结合。这将迫使这两个序列凑得很近，使它们能随后被选择性因子 1（SL1；四个亚基）结合。这种稳定结构使其他因子（未显示）以及随后的 RNA 聚合酶 I 得以结合。

由 RNA 聚合酶Ⅲ进行的转录

RNA 聚合酶Ⅲ亦参与多种持家基因的转录，这些基因编码各种小而稳定的 RNA 分子如 5S rRNA、tRNA 分子、7SL RNA 及某些 RNA 剪接所需的 snRNA 分子。这些基因的特点是启动子位于基因的编码序列中，而不是它的上游。在 tRNA 基因中发现了一种**两分启动子**（bipartite promoter），由两个良好保守的序列，即 A 盒和 B 盒组成，但 5S rRNA 基因的启动子却只含有一个元件，C 盒。在各种情况下，由 RNA 聚合酶Ⅲ进行的转录被认为起始于泛在的转录因子与启动子元件的结合，之后是其他因子的相继结合以及聚合酶最终的聚集（图 10.4）。

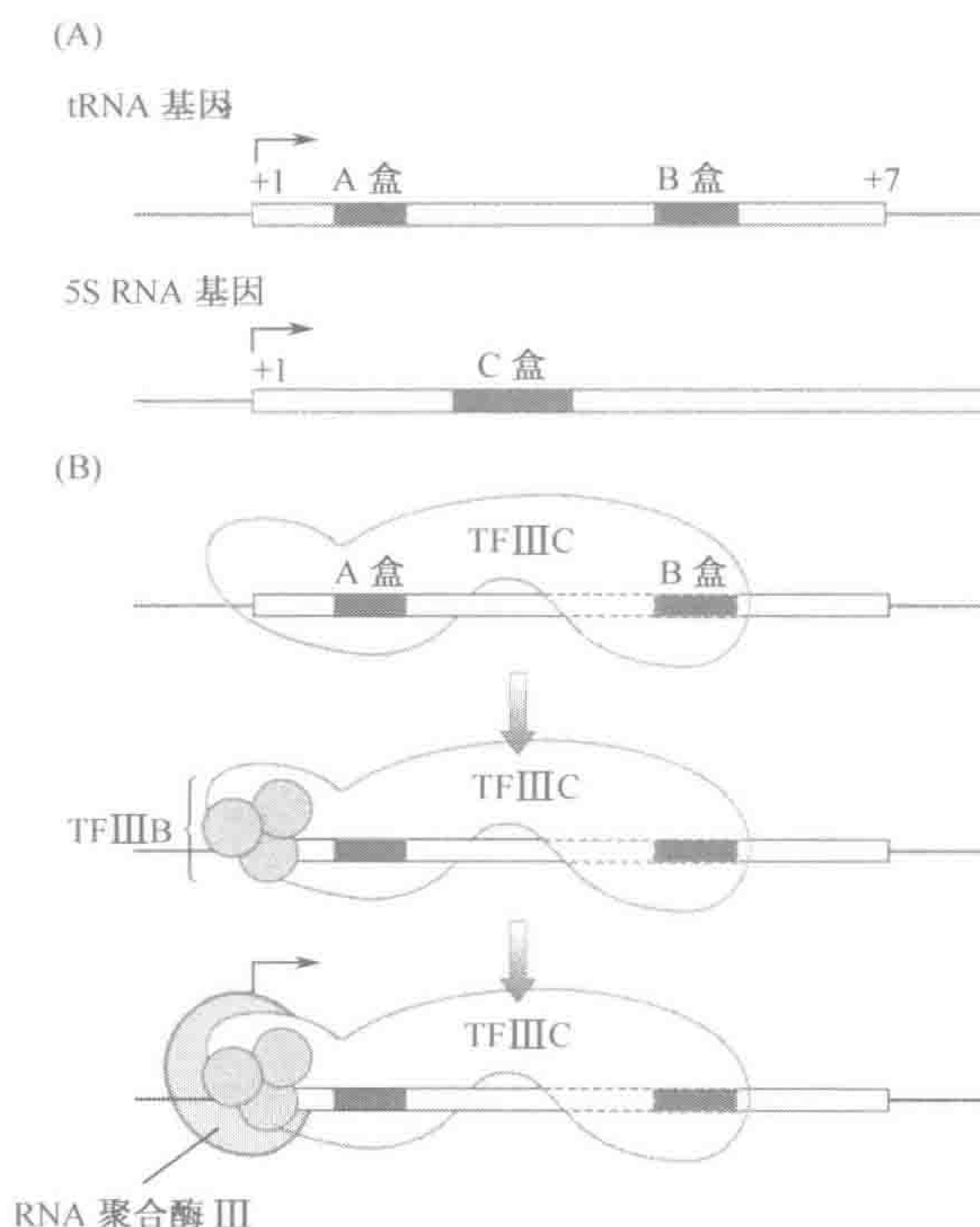


图 10.4　tRNA 和 5S rRNA 具有位于编码序列内的启动子

（A）**tRNA 与 5S rRNA 基因启动子元件的位置**。tRNA 基因内的启动子元件 A 和 B 分别位于编码 D 环和 TψCG 环的序列中（图 1.7B 中的 tRNA 结构）。（B）**tRNA 基因转录的起始**。TFⅢC 转录因子结合至启动子元件使三聚体 TFⅢB 因子能够随后与紧邻转录起始点的上游序列相结合。响应于 TFⅢB 因子的结合，RNA 聚合酶Ⅲ将结合并启始转录。对于 5S rRNA 基因而言，存在一种类似的机制，但另外还需要一个转录因子 TFⅢA 结合至 C 盒，结合的 TFⅢA 因子使 TFⅢB 能够随后结合，随后是 TFⅢC 和 RNA 聚合酶Ⅲ的聚集，如同 tRNA 基因一样。

10.2.3　由 RNA 聚合酶Ⅱ进行的转录需要成套复杂的顺式调节序列和组织特异性转录因子

RNA 聚合酶Ⅱ负责转录所有编码多肽的基因以及特定种类的 snRNA 基因。RNA 聚合酶Ⅱ（同 RNA 聚合酶Ⅰ和Ⅲ一样）依赖于辅助性的通用转录因子，已知的几种可

能具有复杂的结构。例如，TFⅡD 即由 TATA **框结合蛋白**（TATA box-binding protein，TBP）（亦发现与 RNA 聚合酶Ⅰ和Ⅲ相关）加上各种 TBP 相关因子或 **TAF 蛋白**（TAF protein）组成（也见表 10.2）。聚合酶与通用转录因子形成的复合体被称作**基础转录装置**（basal transcription apparatus），并且是起始转录的全部所需。基因将以由核心启动子（如下）决定的最低速度进行组成性表达，除非转录的速度被其他正向或负向调控元件（可能位于一段距离之外或者是启动子区域本身固有的组分）增加或关闭。

表 10.2　真核细胞 RNA 聚合酶Ⅱ和相关通用转录因子的亚基组成

因子	亚基数目	功能
聚合酶Ⅱ	12	催化 RNA 合成
TFⅡD	12	识别核心启动子并提供其他通用转录因子聚集的支架。包括 TATA 结合蛋白(TBP)和几种 TBP 相关因子(TAF)
TFⅡB	1	结合 TBP,选择起始位点并聚集聚合酶Ⅱ
TFⅡA	3	稳定 TFⅡB 与 TBP 的结合
TFⅡF	2	结合聚合酶Ⅱ和 TFⅡB
TFⅡE	2	聚集含有一个解旋酶与一个蛋白激酶的 TFⅡH
TFⅡH	9	在启动子位置解开 DNA 双螺旋;磷酸化聚合酶Ⅱ的 C 端区域,从而造成构象改变。激活的聚合酶自通用转录因子脱离并俘获新的蛋白质以帮助它进行长距离转录而不脱离 DNA

注：通用转录因子的命名使用共同的前缀 TF(transcription factor，转录因子)，接下来以一个罗马数字来表示相关的 RAN 聚合酶。

一些编码多肽的基因是持家基因，但与 RNA 聚合酶Ⅰ和Ⅲ所转录的基因不同，较大比例由 RNA 聚合酶Ⅱ转录的基因呈现组织局限性或组织特异性表达模式。由于同一个体的不同有核细胞中的 DNA 基本一致，一个细胞的特征，不管是肝细胞还是例如 T 淋巴细胞，在很大程度上是由细胞所产生的蛋白质来决定的，因此，除了通用的泛在转录因子外，组织特异性或组织局限性转录因子将通过识别并结合特异性顺式作用序列元件来调节许多编码多肽基因的表达。

部分是由于哺乳动物巨大的核基因组并且也由于拥有极大数量相互作用的基因而造成的对于更为复杂的调控系统的需要，真核细胞中的调控元件甚为精细。单个人类基因的表达经常是由几组顺式作用的调节元件所控制。虽然单个调节元件可能由分布于几百个碱基对范围内的多个短序列元件组成（通常 4～8 个核苷酸），调节一个基因表达的不同类型调控元件可能相隔相当长的距离。多种不同类型的顺式作用元件可被识别，包括启动子、增强子、沉默子、边界元件（绝缘子）以及反应元件等（框 10.2 及图 10.5、10.6）。

框 10.2　参与调节多肽编码基因转录的顺式作用序列元件的种类

启动子（promoter）是用于启始转录的短序列元件的组合（通常位于紧邻基因的上游区域——常常距转录起始点 200 bp 以内）。它们可被细分为不同的组分。

▶ **核心启动子**（core promoter）将指引基础转录复合体以起始基因的转录。在缺乏其他调节元件的情况下，它将容许基因的组成性表达，但仅以很低的（基础）水平。核心启动子元件通常距

框 10.2 参与调节编码多肽基因转录的顺式作用序列元件的种类（续）

转录起始点很近，大约位于−45～+40 核苷酸的位置（Butler and Kadonaga，2002）。如图 10.5 所示，它们包括：位于约−25，为富含 GC 的序列所包围，并且为 TFⅡD 的 TATA 结合蛋白亚基所识别的 TATA 框；紧邻 TATA 元件的上游，位于大约−35，为 TFⅡB 的组分所识别的 BRE 序列；位于转录起始点，并且为 TFⅡD 所结合的 Inr（**起始因子**）**序列**；位于相对转录起始点+30 的位置，并且为 TFⅡD 所识别的 DPE 或**下游启动子元件**（Downstream Promoter Element）等。

▶ **近侧启动子区域**（the proximal promoter region）为紧邻核心启动子上游的序列，通常为−50～−200 bp（更上游的启动子元件将被描述为定位于远侧启动子区域）。其他的**非核心启动子元件**（noncore promoter element）通常位于近侧启动子区域（尽管它们亦可见于核心启动子区）。它们包括：GC 框（又称 Sp1 框，一致性序列为 GGGCGG，后者在转录起始点 100 bp 内可见多个拷贝，并为泛在的 Sp1 转录因子所结合）；CCAAT 框通常位于−75 的位置。CCAAT 框为 CTF（CCAAT 结合转录因子，又称核因子 1，NF-1）以及 CBF（CCAAT 框结合因子，又称核因子 Y，NF-Y）所识别。注意 CCAAT 与 GC 框被用于调节核心启动子的基础转录，并作为增强子序列（见下一节）而发挥作用，而沉默子元件（下面）也可以是启动子整体的组分。

增强子（enhancer）为正向转录调控元件，尤其常见于复杂真核生物诸如哺乳动物等的细胞内，但在简单真核生物如酵母菌中则不存在或极少见（Mertin，2001）。它们被用来增强由核心启动子元件启始的基础转录。它们的功能与核心启动子不同，并不依赖于它们的方向以及在某种程度上与它们所调节基因的距离。增强子元件可位于所调节基因的很远处（例如 β 珠蛋白的基因座调控区，图 10.23）。增强子通常仅包含跨越约 200～300 bp，几种为泛在转录因子所识别的元件加上几种为组织特异性转录因子所识别者（图 10.7）。另外，一些增强子元件可以是启动子整体的组分，如同 CCAAT 与 GC 框（见上文）。

沉默子（silencer）用于降低转录水平。尽管研究得不甚透彻，已鉴定出两类沉默子：**经典沉默子**（classical silencer）（又称沉默子元件）为指引一种主动性转录抑制机制的非位置依赖性元件；**负调控元件**（negative regulatory element）为导致一种被动抑制机制的位置依赖性元件（Ogbourne and Antalis，1998）。在已研究的人类基因中，沉默子元件已被发现于各种位置：邻近启动子、上游某处，还有内含子内部。然而，这类序列的证据常来源于体外 DNA 结合的研究，而它们在活体中的意义仍不确定。

边界元件（boundary element）（绝缘子）（insulator）为常常跨越 0.5 kb～3 kb 的 DNA 区域，其功能是阻断（或隔绝）对转录具有正效应（增强子）或负效应（沉默子，异染色质样抑制效应）的因子影响的传播（Bell *et al.*，2001）。

反应元件（response element）响应于特定的外界刺激而调节转录。它们通常位于启动子元件上游不远处（通常距转录起始点 1 kb 以内）。多种的这类元件对特定的激素（如视黄醇或类固醇激素诸如糖皮质激素）或细胞内第二信使诸如环化 AMP（节 10.2.5 和表 10.4）具有反应。

基因表达的组织特异性与发育阶段特异性常常由增强子和沉默子序列赋予，已发现多种顺式作用序列可被**组织特异性转录因子**（tissue-specific transcription factor）特异识别。例如，红细胞系统中的特异性表达经常由两种序列之一来给予信号：TGACTCAG（或其反向互补物 CTGAGTCA：二者均为红细胞系统特异性转录因子 NF-E2 所识别）；或者序列 [A/T]GATA[A/G]（或其反向互补物，二者均为 GATA 系列的转录因子所识别。见图 10.7 的例子，以及表 10.3 中的其他组织特异性顺式作用元件）。

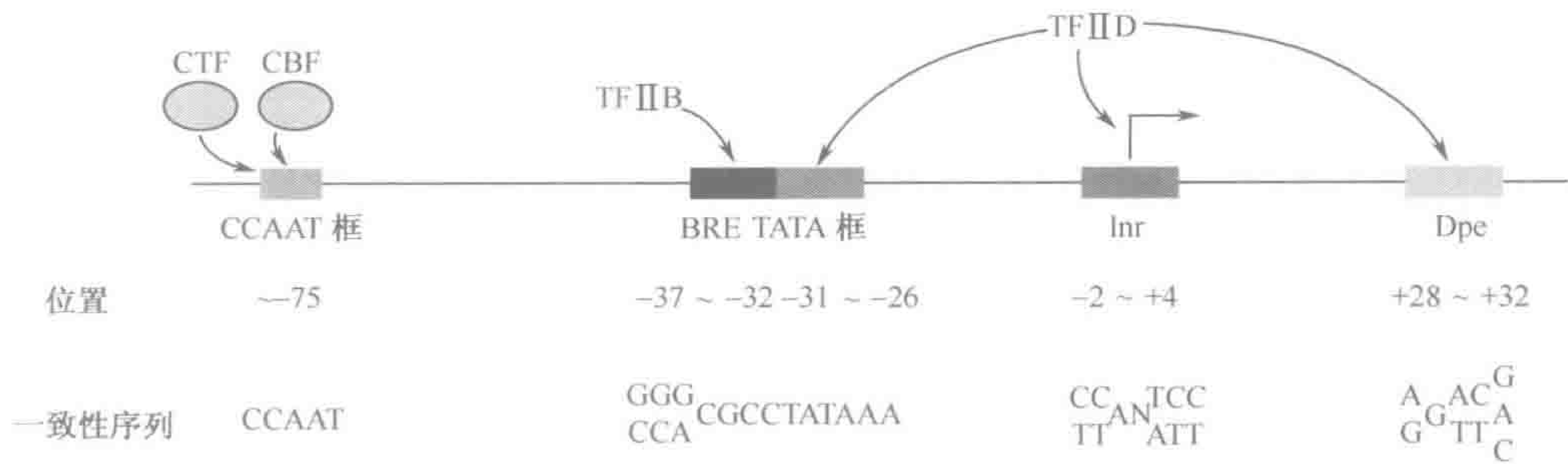


图 10.5　复杂真核生物中为泛在转录因子所结合的调节启动子元件的保守位置

注：单个基因的核心启动子无需包含所有元件。例如，许多启动子缺乏 TATA 框，而是使用功能相近的起始因子（INR）元件。GC 框亦常见于启动子中，但它们的位置更为多变（部分例子见图 10.6 和图 1.13）。BRE，TFⅡB 识别元件；DPE，下游启动子元件。经冷泉港实验室杂志允许，改编自 Butler 和 Kadonga(2002). Genes Dev，16，2583～2592。

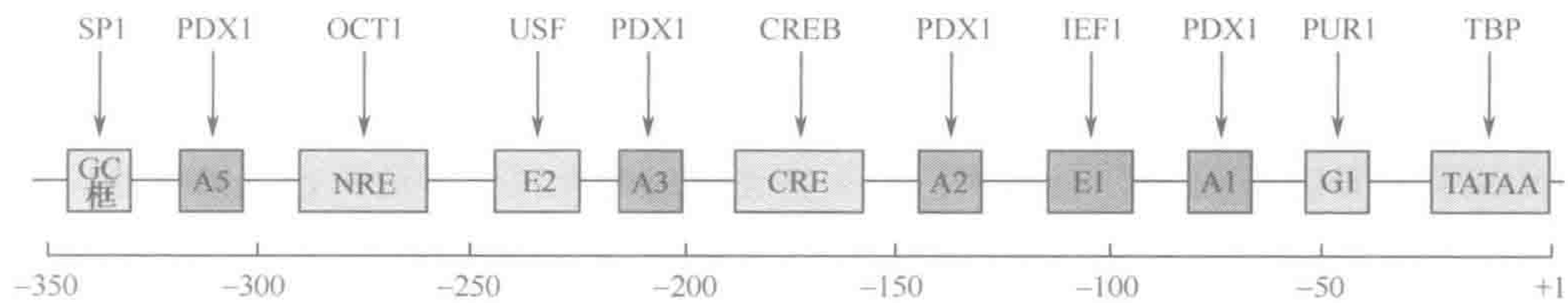


图 10.6　人类胰岛素基因启动子包含多种为泛在及组织特异性转录因子所识别的序列元件

箭头指示转录因子（上排）与存在于人类胰岛素基因上游的调节序列元件（框内）的结合。泛在或广泛表达的转录因子用浅灰色表示；用浅灰色表示者为胰腺 β 细胞特异性。PDX1 转录因子与存在于胰岛素基因启动子内的四个 C(C/T)TAATG 形式的序列基序（A1、A2、A3、A5）相结合。缩写：CRE，cAMP 反应元件；NRE，负调控元件。

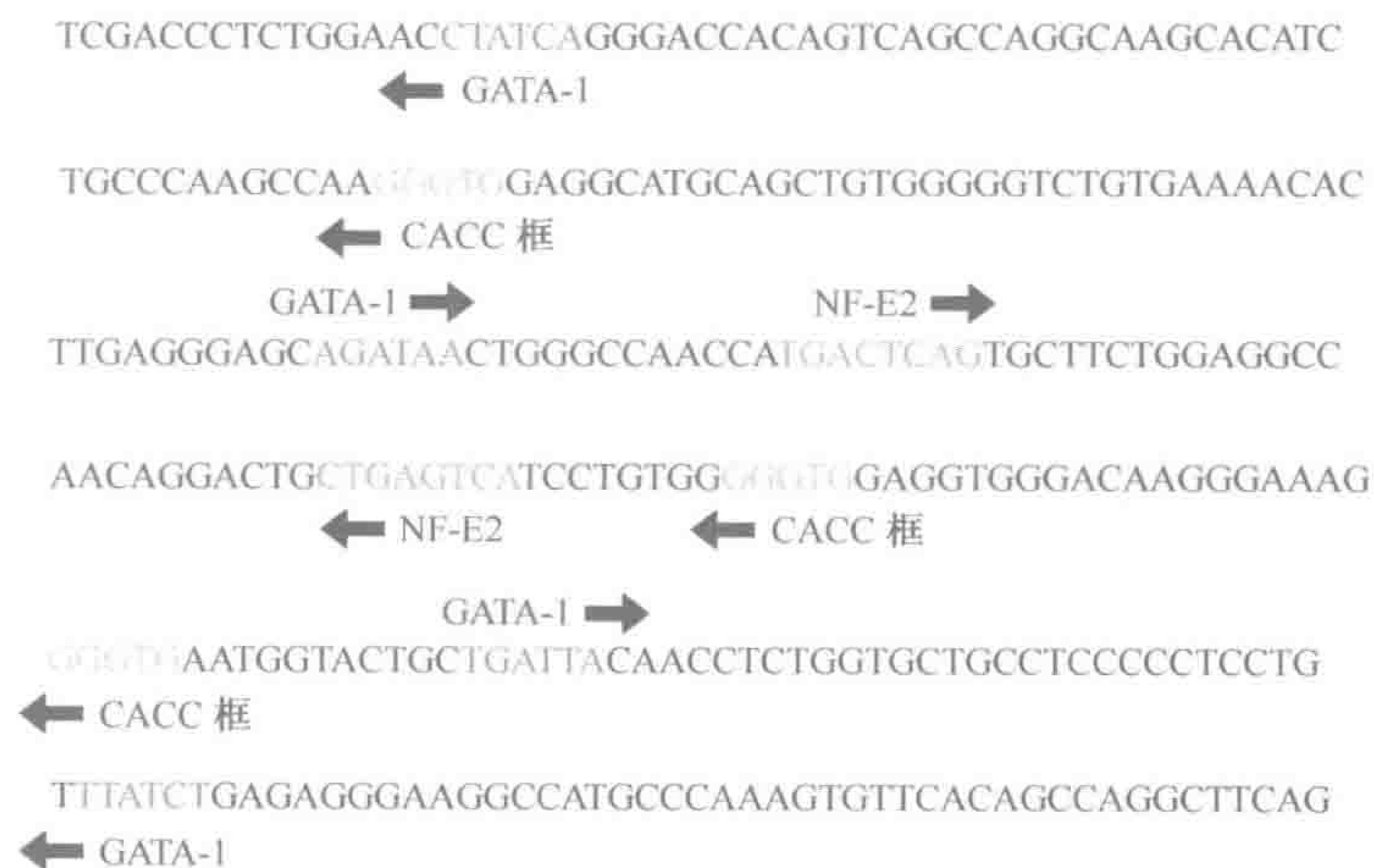


图 10.7　HS-40 α 珠蛋白基因调节位点含有许多红细胞系统特异性转录因子的识别元件

注：HS-40 位点似乎是 α 珠蛋白基因簇（节 10.5.2）的一个基因座调控区。

表 10.3　为组织局限性和组织特异性转录因子所识别的顺式作用序列的例子

一致性结合序列	转录因子	表达模式
(A/T)GATA(A/G)	GATA-1、-2 等	红细胞系统
TGACTCAG	NF-E2	红细胞系统
GTTAATNATTAAC(=PE 元件)	HNF-1	肝脏、肾脏、胃、肠、脾
T(G/A)TTG(C/T)	HNF-5	肝脏
ATGCAAT	POU2F2(OTF-2)	淋巴样细胞
GCCTGCAGGC	Ker1	角质细胞
(C/T)TAAAAATAA(C/T)3	MBF-1	肌细胞
(C/T)TA(A/T)AAATA(A/G)	MBF-2	肌细胞
CAACTGAC	MyoD	成肌细胞+肌管
(C/A)A(C/A)AG	TCF-1	T 细胞

除活跃地促进组织特异性转录以外，部分顺式作用的沉默子元件通过阻断除所需组织外所有组织中的表达而赋予组织或发育阶段的特异性。例如，神经局限性沉默子元件(NRSE)将抑制除神经组织外所有组织中若干基因的表达（Schoenherr *et al*.,1996）。一种与 NRSE 结合的转录因子，被各异地称为 NRSF（神经局限性沉默因子）或 REST（RE-1 沉默转录因子），其广泛地表达于非神经组织和发育早期的神经元前体中，但随后它将特异性地不表达于更为成熟的（有丝分裂期后的）神经元中。NRSF/REST 似乎能够聚集一种辅阻遏物 co-REST，后者似乎用以为聚集的分子机构，该机构能够诱导横跨含有神经元的表达基因的染色体区域的转录沉默（Lunyak *et al*.,2002）。

10.2.4　转录因子含有容许 DNA 结合的保守性结构基序

转录因子识别并结合一段短的核苷酸序列，通常是由于蛋白质表面与所结合区域双螺旋的表面特征之间的广泛互补性。尽管氨基酸与核苷酸之间的个别相互作用很微弱（通常为氢键、离子键以及疏水作用），但 DNA-蛋白质结合的区域则通常以大约 20 个这类作用为特征，后者将共同确保这种结合很强且特异。在人类及其他真核转录因子中，两种独特的功能常常能被发现并定位于其蛋白质的不同位置上：

- **激活结构域**（activation domain）一旦有转录因子与其结合，即可激活靶基因的转录。激活结构域被认为通过与基础转录因子相互作用来刺激转录，从而有助于在启动子上形成转录复合体。尽管不如 DNA 结合结构域研究得透彻，但其中一些据知富含天冬氨酸和谷氨酸残基（酸性激活结构域）；另一些则富含脯氨酸或谷氨酸。
- **DNA 结合结构域**（DNA-binding domain）将容许转录因子与靶基因特异性地结合。在许多不同的转录因子中已发现了若干普遍性的保守性结构基序，包括下面将要详述的亮氨酸拉链、螺旋-环-螺旋、螺旋-转角-螺旋以及锌指基序等。每种基序都通过 α 螺旋（或者偶尔为 β 折叠，见图 1.24）结合至 DNA 的大沟。显而易见，尽管这些基序在大体上提供了 DNA 结合的基础，但 DNA 结合域中的序列元件的精确聚集则将为所需的序列特异性识别提供基础。大多数转录因子以同型二聚体的形式结合到 DNA 上，而蛋白质的 DNA 结合区域则通常不同于负责形成二聚体的区域。

亮氨酸拉链基序（leucine zipper motif）

亮氨酸拉链为富含亮氨酸残基的螺旋状伸展的氨基酸（通常每七个氨基酸、即每两圈螺旋即出现一次，见图 10.8），后者易于形成一种二聚体。每个单体单位由一个两性的 α 螺旋（组成氨基酸的疏水侧基朝向一面；极性基团朝向另一面，图 1.24）组成。单个单体的两个 α 螺旋跨越很近的距离连接形成一个卷曲的盘状结构（节 1.5.5），以主要的相互作用发生于个体单体两个相反的疏水氨基酸之间。在这个区域以外，两个 α 螺旋将相互分离，从而使整个二聚体呈 Y 形结构。这种二聚体被认为能够拉紧双螺旋，很像衣服的拉链锁住一条衣缝（图 10.9）。除形成同型二聚体之外，亮氨酸拉链蛋白偶尔也能依靠两个不同单体的疏水表面的相容性形成杂二聚体。这种杂二聚体的形成为基因调控提供了一种重要的组合调控机制。

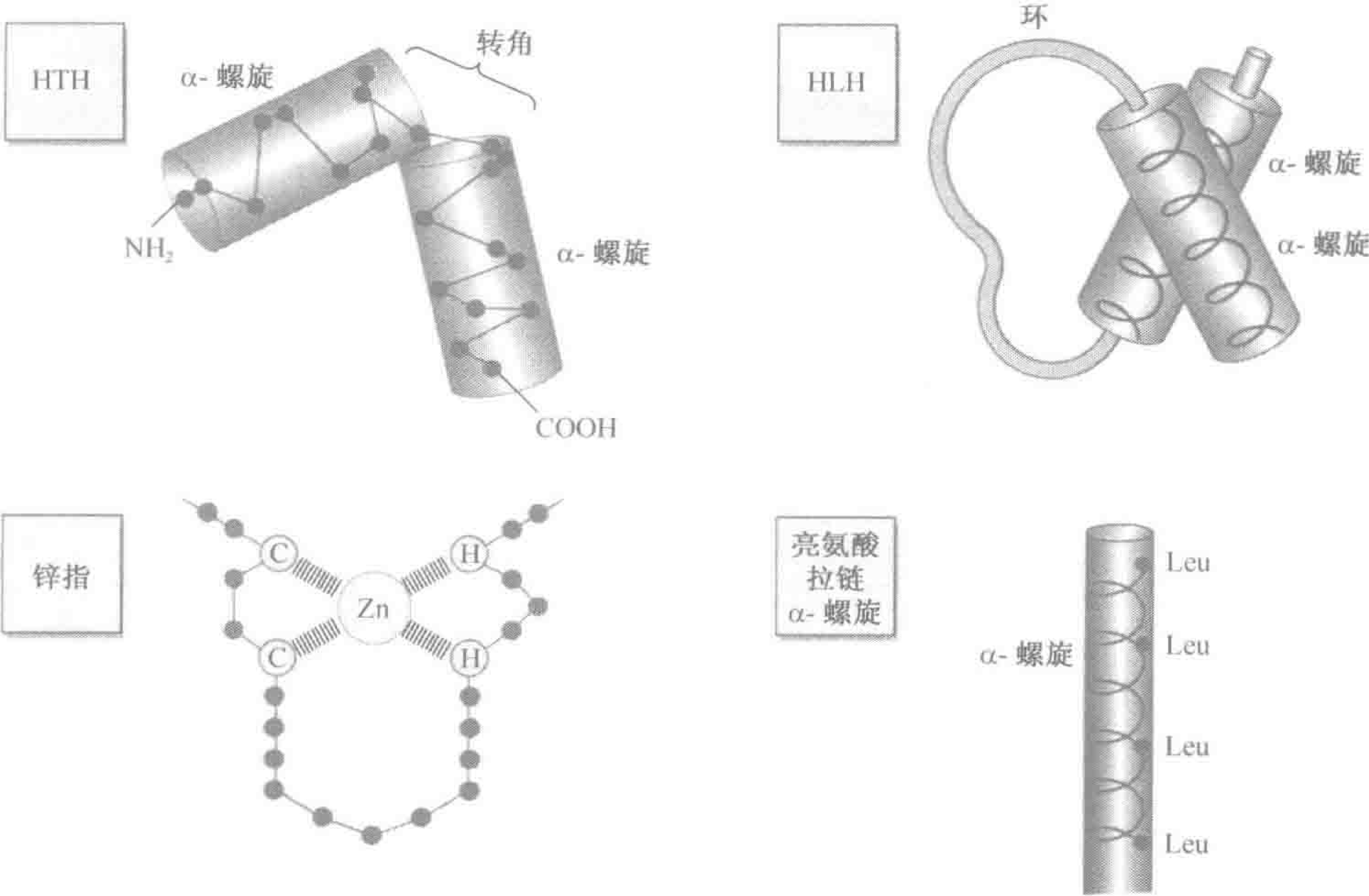


图 10.8　常见于转录因子与 DNA 结合蛋白中的结构基序

缩写：HTH，螺旋-转角-螺旋；HLH，螺旋-环-螺旋。**注**：亮氨酸拉链单体为两性分子［即具有疏水残基（亮氨酸）始终如一地出现于螺旋的一侧，见图 1.24］。两个这样的螺旋能够排成一行，以它们的疏水面相对而形成一个卷曲的盘状结构。

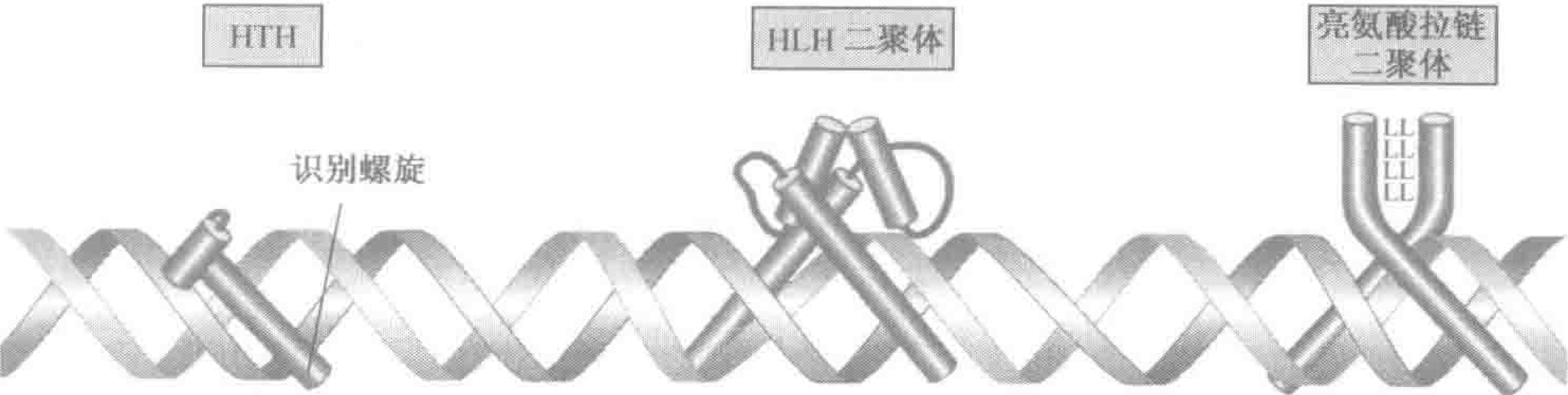


图 10.9　转录因子中的保守结构基序与双螺旋的结合

注：螺旋-环-螺旋（HLH）二聚体与亮氨酸拉链二聚体的单个单体以不同的颜色表示以供区别，但也可能完全相同（同型二聚体）。HLH 杂二聚体和亮氨酸拉链杂二聚体可能提供更高水平上的调控（见正文）。

螺旋-环-螺旋基序（helix-loop-helix motif）

螺旋-环-螺旋（HLH）基序与亮氨酸拉链相关，并需要与下节描述的螺旋-转角-螺旋（HTH）基序相区别。它由两个α螺旋，一长一短，通过一个可折叠的环相连。与HTH基序中的短转角不同，HLH基序中的环具有足够的可折叠性以允许折回，因此两个螺旋可彼此压紧，即两个螺旋将位于相互平行的平面上，与HTH基序中的两个螺旋形成对比（图10.8）。HLH基序将介导DNA结合以及蛋白质二聚体的形成（图10.9），并且它容许偶尔的异二聚体形成。然而，在后一种情况中，异二聚体将形成于一个全长HLH蛋白与一个缺乏结合DNA所必需的全长α螺旋的截短的HLH蛋白之间。所产生的HLH异二聚体无法与DNA紧密结合。因此，HLH杂二聚体被认为通过使特定的基因调控蛋白失活而形成一种调控机制。

螺旋-转角-螺旋基序（helix-turn-helix motif）

HTH基序为发现于同源框和一些其他转录因子中的常见基序。它包括两个短的α螺旋，被一个短的氨基酸序列分隔，后者将导致一个转角，因此两个α螺旋将具有不同的方向（即两个α螺旋不在一个平面，与HLH基序中的情况不同；图10.8）。这种结构与几种噬菌体调控蛋白的DNA结合基序极为相似，如λcro蛋白，它与DNA的结合已通过X射线晶体学进行了深入的研究。对λcro蛋白和真核细胞的HTH基序而言，看法是虽然HLH基序在大体上将介导DNA结合，更靠近C端的螺旋将作为一种特异性的**识别螺旋**（recognition helix），因为它与DNA的大沟相匹配（图10.9），从而精确调控所识别的DNA序列。

锌指基序（zinc finger motif）

锌指基序涉及四个保守的氨基酸与一个锌离子的结合，以形成一个环（指），一种经常为串连重复的结构。尽管存在几种不同的形式，但常见的形式将涉及由两个保守的半胱氨酸残基和两个保守的组氨酸残基，或者四个保守的半胱氨酸残基与一个Zn^{2+}离子的结合。所形成的结构因此可能包括通过Zn^{2+}离子协调聚集在一起的一个α螺旋和一个β折叠，或者是两个α螺旋。无论如何，与DNA的主要接触是通过一个α螺旋与大沟的结合。所谓的C2H2(Cys_2/His_2）锌指通常由23个氨基酸组成，与相邻的指被大约七个或八个氨基酸分开（图10.8）。

10.2.5 容许对外部刺激产生反应的基因表达转录调节的多种机制

在真核细胞中，基因表达能够在当细胞分化时以一种半永久的方式，或者反应于细胞外信号而以一种暂时、易于逆转的方式（**可诱导基因表达**，inducible gene expression）发生改变。诸如细胞外特定离子和小营养分子的浓度、温度刺激等环境因素均可导致暴露于这些参量变化的细胞内基因表达模式的显著改变。在复杂得多细胞动物中，亦存在对于细胞间相互交流的基本需求，而各种方式的**细胞信号**（cell signaling）均存在可能（节3.2.1，3.2.2）。在某些情况下，基因表达的改变表现在翻译水平，后者可能带来一定的优势。在另一些情况下，基因表达是通过对转录的调节来改变的。

反应于细胞信号的转录调节可能采用不同的形式，但其终点则总是相同的：原先失活的转录因子被特异性地激活，并于随后与位于靶基因启动子内的特异性调节序列相结合，从而调节它们的转录。对于通过信号分子或它们的中介物所调节的转录而言，这类调节序列通常被称为**反应元件**（response element）（表 10.4）。

表 10.4 可诱导基因表达中反应元件的例子

一致性反应元件(R. E.)	响应于	识别 R. E. 的蛋白因子
(T/G)(T/A)CGTCA	cAMP	CREB(又称 ATF)
CC(A/T)(A/T)(A/T)(A/T)(A/T)GG	血清生长因子	血清反应因子
TTNCNNNAAA	γ 干扰素	Stat-1
TGCGCCCGCC	重金属	Mep-1
TGAGTCAG	巴豆油酯	AP1
CTNGAATNTTCTAGA	热休克	HSP70 等

注：另外参见图 10.10 中的激素反应元件。

配体诱导性转录因子

小分子的疏水性激素和诸如类固醇激素、甲状腺素以及视黄酸等成形素能够通过靶细胞的质膜扩散并结合至细胞质或核内的受体上。这些受体［通常称为**激素核受体**（hormone nuclear receptor）］为可诱导的转录因子：在同源配体结合之后，受体蛋白将与位于大概 50～100 个靶基因启动子区域中的一个特定 DNA 反应元件相联合，并激活它们的转录。

尽管甲状腺素和视黄酸在结构及生物合成上与类固醇激素无关，它们的受体却同属一个常见的核受体超家族。两个保守性结构域是该家族的特性：一个位置居中的约 68 个氨基酸的 **DNA 结合结构域**（DNA-binding domain）以及一个位于近 C 端的约 240 个氨基酸的**配体结合结构域**（ligand-binding domain）（图 10.10）。DNA 结合结构域含有锌指结构，以二聚体的形式结合，每个单体识别反应元件中两个六核苷酸之一。这两个六核苷酸或为反向重复，或为同向重复，通常被三个或五个核苷酸分开（图 10.10）。当缺乏配体时，受体将通过配体结合结构域直接抑制 DNA 结合结构域的功能或结合至一个抑制蛋白而失活，如同糖皮质激素受体的情况（图 10.11）。

由信号转导所致的转录因子的激活

与脂溶性激素或成形素不同，亲水的信号分子诸如多肽激素等不能扩散通过浆膜。相反，它们与细胞表面的特异性受体结合。与配体分子结合之后，受体将经历某种构象改变，并以通过细胞内的其他分子将信号传递下去的方式被激活（**信号转导**，signal transduction；见节 3.2.1）。

许多细胞表面受体具有激酶活性或能够激活细胞内的激酶（表 3.3），信号转导途径通常是以激酶和磷酸酶之间复杂的调节性相互作用为特征，磷酸酶能够通过磷酸化/去磷酸化而激活或抑制中介物。在许多情况下，磷酸化或去磷酸化将诱导构象的改变。对于信号分子的激活而言，改变了的构象常常意味着信号因子不再被它所结合的抑制蛋

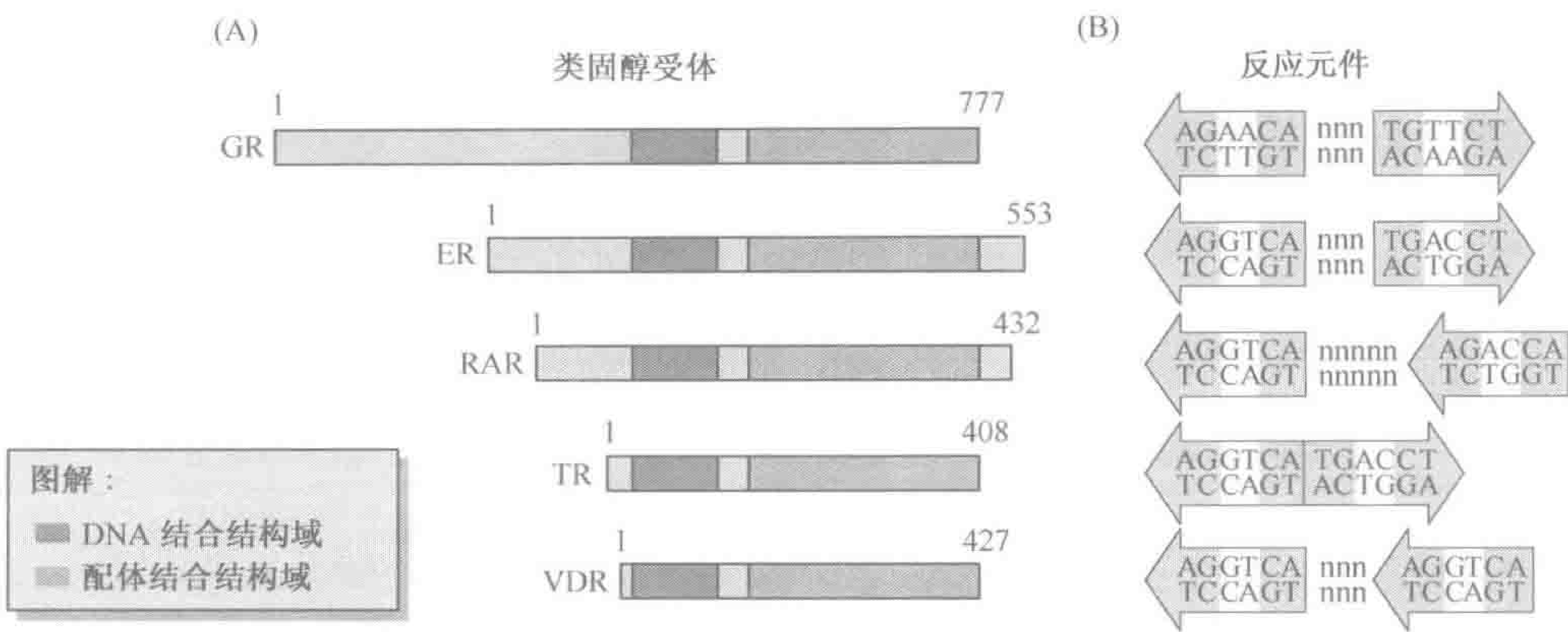


图 10.10　类固醇受体以及各自的反应元件

(A) **核受体超家族成员的结构**。数字指用氨基酸数衡量的蛋白质大小。ER，雌激素受体；GR，糖皮质激素受体；PR，孕激素受体；RAR，视黄醇受体；TR，甲状腺素受体；VDR，维生素 D 受体。 (B) **反应元件** (response element)。**注**：(Ⅰ) 反应元件常常为完全的六核苷酸反向重复，但视黄醇和维生素 D_3 的反应元件则是不完全的六核苷酸同向重复；(Ⅱ) 六核苷酸均具有通用的 AGNNCA 序列，中间的两个核苷酸（白底所示）赋予特异性，并属于以下三种之一：GT、AC 或 AA。

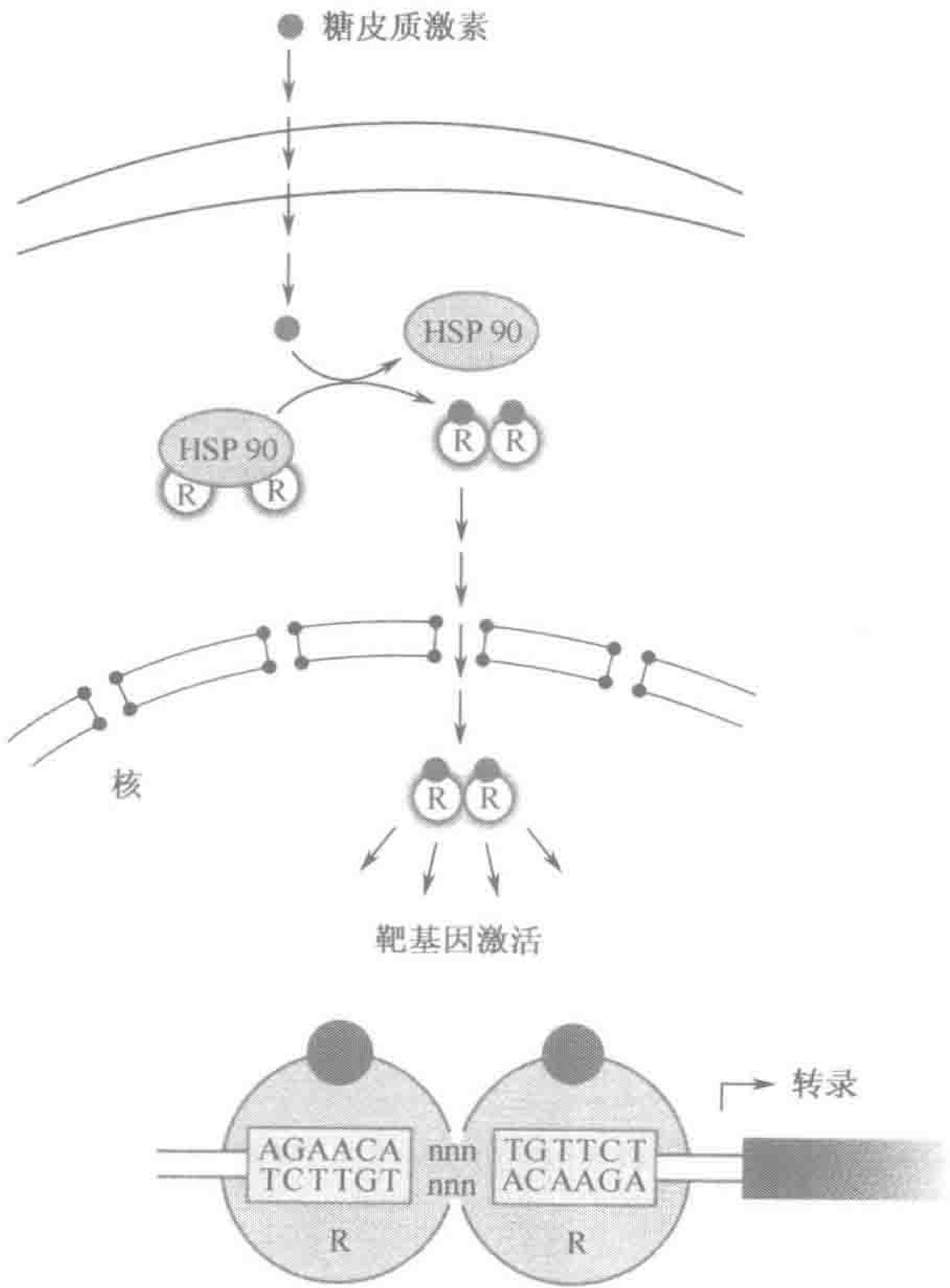


图 10.11　糖皮质激素对转录的调节

糖皮质激素受体在正常情况下将被一个抑制蛋白 Hsp90 结合而失活。糖皮质激素与糖皮质激素受体的结合将释放 Hsp90，受体将二聚化并于之后激活启动子中具有糖皮质激素反应元件的特定基因（图 10.10）。

白中存在的一些抑制因子序列，或者被其自身结构中的某个结构域或序列基序所抑制。

在转录激活方面，两种普遍机制将容许信号自细胞表面受体快速传递至细胞核，二者均涉及蛋白质磷酸化：蛋白激酶被激活并随后从细胞质转运至细胞核，在那里它们将磷酸化目标转录因子；储藏于细胞质中的非活性转录因子将通过磷酸化激活并转移至细胞核内。下面两节将举例说明上述两种机制（另见 Karin and Hunter，1995）。

通过环化 AMP 通路的激素信号

环化 AMP 为一种对多种激素和其他信号分子发生反应的重要**第二信使**（second messenger）（表 10.5）。它由一种膜结合酶，即腺苷酸环化酶自 ATP 合成。激活腺苷酸环化酶的激素与细胞表面的一种 G 蛋白偶联受体类型的受体结合。激素与受体的结合将促进该受体与一种由 α、β 和 γ 三个亚基组成的 G 蛋白之间的相互作用。在这种相互作用下，G 蛋白的 α 亚基被激活，造成它的分离并刺激腺苷酸环化酶。

表 10.5　细胞信号中第二信使的例子

第二信使	特征
环化 AMP(cAMP)	由腺苷酸环化酶自 ATP 产生。作用通常由蛋白激酶 A 介导。见 CREB 因子激活的例子(图 10.12A)。
环化 GMP(cGMP)	由鸟苷酸环化酶自 GTP 产生。最具特征的作用是脊椎动物眼睛的视觉接收。
磷脂/Ca^{2+}	在 G 蛋白偶联受体及蛋白质酪氨酸激酶的下游激活。4,5-二磷酸磷脂酰肌醇(PIP_2)水解将产生甘油二酯以及激活蛋白激酶 C 并聚集细胞内存储的 Ca^{2+} 的 1,4,5-三磷酸肌醇(IP_3)。

由激活的腺苷酸环化酶所致的细胞内 cAMP 的增加随后将激活含有特异性 cAMP 反应元件或称 CRE 的特异性靶序列的转录。cAMP 的这种功能由蛋白激酶 A 介导。通过容许释放两个具有催化活性的亚基，后者随后进入细胞核并磷酸化一种特定的转录因子 CREB（CRE 结合蛋白），环化 AMP 与蛋白激酶结合并将其激活。活化的 CREB 随后将激活含有 cAMP 反应元件的基因的转录（图 10.12A）。

由肿瘤坏死因子信号激活 NF-κB

NF-κB 为一种涉及多种方面免疫反应的转录因子。在非活性状态下，NF-κB 被保留于细胞质中，并与一种抑制亚基 IκB 形成复合体。然而，后者可在磷酸化作用下成为降解的目标。随之发生的 IκB 的破坏将容许 NF-κB 转移至细胞核并激活它的各种靶基因。在肿瘤坏死因子（TNF）与特异性的细胞表面 TNF 受体结合并导致 TNF 受体相关因子 2（TRAF2，见图 10.12B）被激活之后，IκB 的磷酸化将由一个激酶级联来实现。

10.2.6　基因表达的翻译调控可涉及 RNA 结合蛋白对 UTR 调节序列的识别

蛋白质合成是基因表达主要的最终步骤，也是调节的一个重要控制点。真核细胞中一系列错综复杂的蛋白质涉及翻译的起始，并且不同的翻译起始途径已被揭示（Dever，2002）。起始密码子 AUG（蛋氨酸）的选择是关键之一，但在一些基因中对 mRNA 中

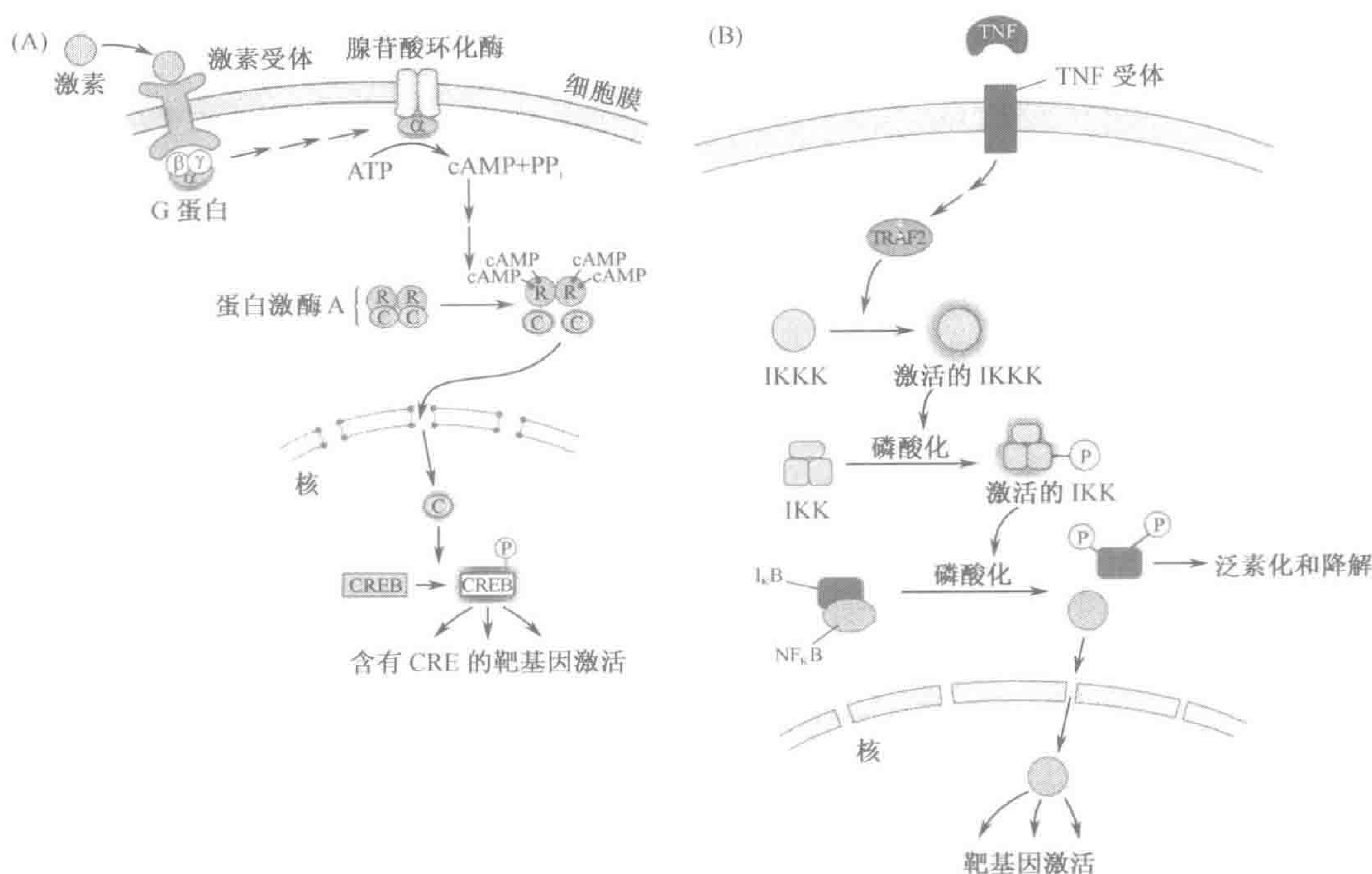


图 10.12　特定的靶基因可通过自激活的细胞表面受体的信号转导对外界刺激发生反应而活跃表达
(A) **蛋白激酶激活并转移至细胞核**：通过环化 AMP-蛋白激酶 A 信号转导通路的激素信号途径。激素与特异性细胞表面受体的结合将促进该受体与 G 蛋白的相互作用。激活的 G 蛋白 α 亚基将从受体分离并刺激膜结合腺苷酸环化酶以合成 cAMP。后者将与蛋白激酶 A 的调控亚基相结合，使之能够释放催化亚基（c），该亚基将转移至细胞核，通过磷酸化激活转录因子 CREB（CRE 结合蛋白）。激活的 CREB 将与靶基因启动子的 cAMP 反应元件相结合。(B) **细胞质转录因子（NF-κB）的激活以及向细胞核的转移**。TNF 的结合将引起其膜受体的改变，容许它聚集许多细胞内的信号蛋白。它们随之将聚集并激活 IKKK，一种磷酸化 IκB 的激酶，即 IκB 激酶。IκB 通常与 NFκB 结合，但 IκB 上添加的磷酸基团使其成为泛素化及降解的目标。释放的 NFκB 具有一个暴露的核定位序列，容许其转移至细胞核并在那里激活一组靶基因。

的起始密码子则将作出其他选择，产生 N 端序列各异的异构体——见图 10.16A 中 Wilms 肿瘤基因 *WT1* 的例子。这些异构体的意义尚不十分清楚。

越来越多的真核的及哺乳动物 mRNA 种类被证实在它们的非翻译序列中含有调控序列，以 3′端最为频繁（Wickens *et al*.,1997）。若干真核的及哺乳动物 RNA 结合蛋白亦已发现，并证实与出现于非翻译序列中的特异性调节序列相结合，因而提供了基因表达翻译调控的基础（Siomi and Dreyfuss，1997）。多种不同的 RNA **结合结构域**（RNA-binding domain）已被识别，而它们包含原先与诸如锌指结构和同源结构域等转录因子的 DNA 结合特性相关联的元件（Siomi and Dreyfuss，1997）。

RNA 运输

RNA 中的顺式作用调节元件与反式作用的 RNA 结合蛋白之间的相互作用可设想通过不同途径来改变 RNA 的结构：协助或阻碍与其他反式作用因子的相互作用；改变 RNA 的高级结构；将最初远离的 RNA 序列凑到一起，或者为 RNA 分子转运至特定细

胞内位置（**RNA 运输**，RNA trafficking）提供定位或靶信号。众多真核的及哺乳动物的 mRNA 据知在一些种类的细胞中，特别是神经系统的细胞中是以核蛋白（RNP）颗粒的形式被转运至特定的位置（Hazelrigg，1998）。例如，tau mRNA 定位于轴突的近端而非树突，而许多 mRNA 被定位于成熟的神经元中，髓磷脂基础蛋白 mRNA 在驱动蛋白的协助下将转运到少突胶质细胞的突起中。

RNA 运输可为定位蛋白质而不是单纯地转运它们提供一种更为有效的途径：假设能够结合核糖体的话，一个 mRNA 能产生多种不同的蛋白质分子。现已设想了不同的相继步骤：最初的翻译抑制、细胞内转运、定位（至特定的亚细胞目的地）以及之后的位置依赖性翻译。最近，这一过程中各个步骤所需的关键调节序列已经在非翻译序列、主要是许多 mRNA 种类的 3′UTR 中被发现（Wickens *et al.*,2002）。

外界刺激反应的基因表达翻译调控

基因表达的翻译调控可容许对变化的环境刺激产生比选择性的转录激活更为迅速的反应。铁代谢提供了两个有用的例子。铁水平的增加将刺激铁结合蛋白、即铁蛋白的合成，而并不伴随铁蛋白 mRNA 数量的任何增加。相反，降低的铁含量将刺激转铁蛋白受体（TFR）的产生，对转铁蛋白受体 mRNA 的产量则没有任何影响。铁蛋白重链 mRNA 与轻链 mRNA 的 5′UTR 中均包含一个**铁反应元件**（IRE），一个可形成发夹结构的特异性顺式作用调节序列。在转铁蛋白受体 mRNA 的 3′UTR 中也发现了几个这样的 IRE 序列（Klausner *et al.*,1993）。调控是通过一个特异性的 IRE 结合蛋白与 IRE 的结合来施加的，这种结合在低铁水平时被激活（图 10.13）。

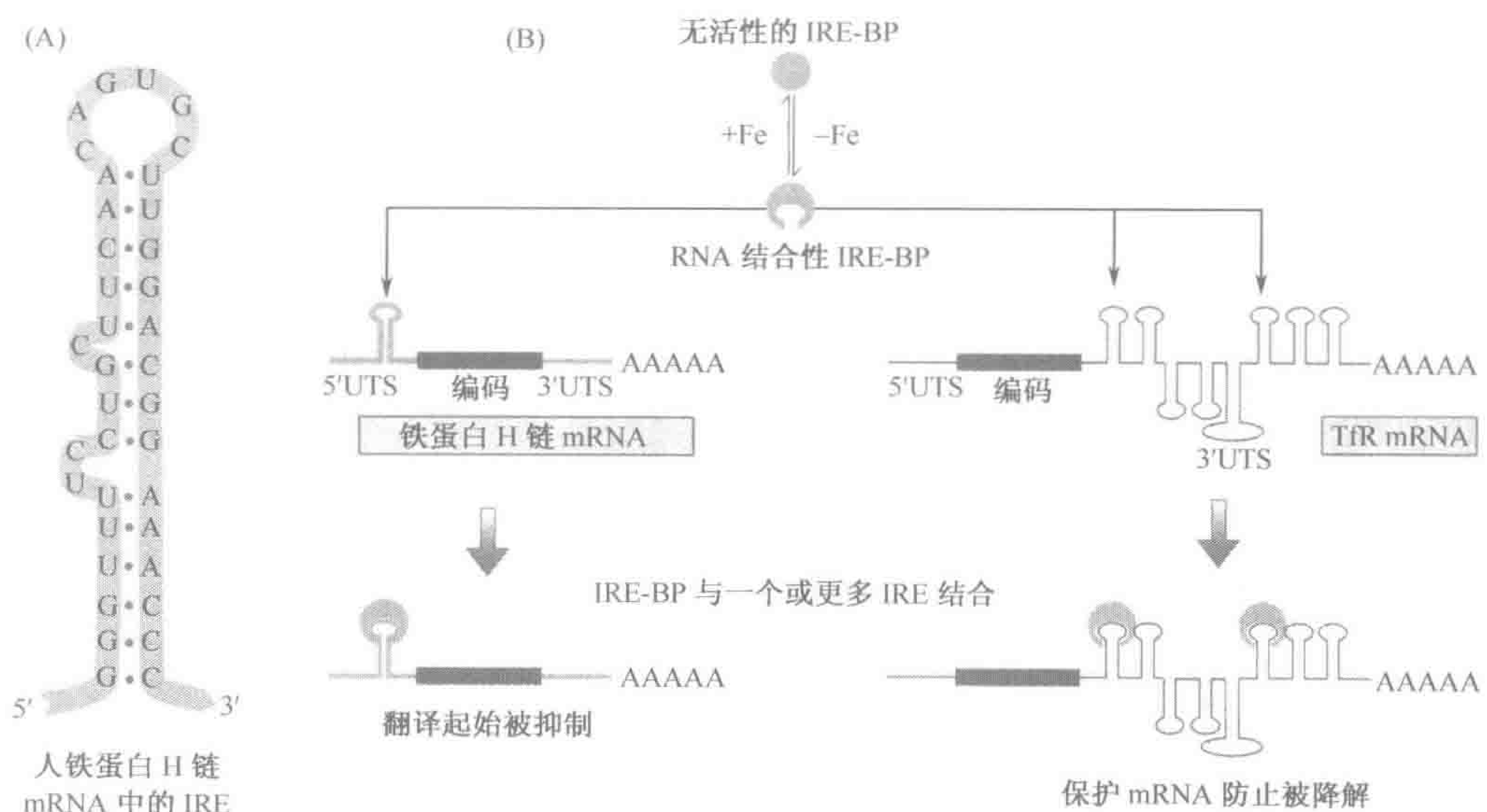


图 10.13　IRE 结合蛋白通过与 5′或 3′非翻译区内的铁反应元件（IRE）相结合而调节铁蛋白重链和转铁蛋白受体的产生

(A) 位于铁蛋白重链 5′UTR 内的 IRE 的结构。(B) IRE 结合蛋白与铁蛋白和转铁蛋白受体 mRNA 的结合对于蛋白质合成具有相反的效应。

发育早期基因表达的翻译调控

在卵母细胞成熟的过程中以及胚胎最早期的基因表达调节于翻译而非转录的水平（De Moor and Richter，2001）。人类卵母细胞受精之后，最初并无 mRNA 产生，直至 4～8 细胞阶段，这时**合子转录**（zygotic transcription）将被激活，即对出现于合子中的基因的转录。在此之前，细胞的功能由原先合成于卵子发生过程中的**母源** mRNA（maternal mRNA）决定。

由模式生物研究的推论认为多种 mRNA 以一种非活性的形式储藏在卵母细胞中，以具有短的 oligo(A) 尾巴为特征。这类 mRNA 之前曾遭受脱腺苷化作用，所产生的短 oligo(A) 尾巴意味着它们不能被翻译。随后，在受精或稍后的发育中，储藏的非活性 mRNA 可通过**细胞质多聚腺苷化**（cytoplasmic polyadenylation）激活，恢复正常大小的 poly(A) 尾巴。相同类型的 poly(A) 聚合酶活性被使用，如同新形成的 mRNA（发生于细胞核中）的标准的多聚腺苷酸化，但除了 AAUAAA 信号外，这种 mRNA 需要有一个富含尿嘧啶的上游**细胞质多聚腺苷酸化元件**（cytoplasmic polyadenylation element）（Wahle and Kuhn，1997）。

另外两种在发育过程中调节部分 mRNA 翻译的机制为**翻译遮蔽**（translational masking）（借此 RNA 结合蛋白可识别并与 mRNA 3′UTR 内的特异性序列相结合，从而抑制翻译；见 Gray and Wickens，1998）和**反义调节**（antisense regulation）。在后者中，一些 mRNA 据知受一种互补 RNA 序列的调节，如同 microRNA，被证实在某些情况下可通过与互补序列的 3′UTR 相结合而在发育过程中调节基因的极小 RNA 的情况（图 9.6 和 Bannerjee and Slack，2002；Pasquinelli and Ruvkun，2002）。

10.3　单个基因的选择性转录与加工

除了通过选择激活或抑制特定的基因（或它们的转录物）来调控外，基因调控机制还可通过选择单个基因特异的不同转录物来实现。启动子的差异性使用和 RNA 差异性加工事件可产生大量不同的异构体，这些和其他机制已经挑战了传统的基因的概念。

10.3.1　选择性启动子的使用可产生组织特异性异构体

若干特殊的哺乳动物基因据知有两个或更多的**选择性启动子**（alternative promoter），后者可产生具有不同特性的选择性表达产物（**异构体**，isoform）（Ayoubi and van de Ven，1996）。通常每个启动子从不同样式的第一外显子驱动转录，后者随后将在各种情况下被剪接至一组共同的下游外显子。然而，另外一些选择性启动子位于基因更远端的部分（**选择性内部启动子**）并驱动截短的蛋白质产物的表达，如同抗肌萎缩蛋白基因中一些启动子的情况（见下文）。这些异构体可提供：

- 组织特异性（一种时常发生的情况，因为不同的启动子可包含不同的调节元件；见下面的人类抗肌萎缩蛋白基因的例子）；
- 发育阶段特异性（例如，胰岛素样生长因子Ⅱ基因）；
- 差异性的亚细胞定位（例如，可溶性与膜结合性异构体）；

▶ 差异性的功能特性（如黄体酮受体的情况）；

▶ 性别特异性基因调节［*Dnmt1* 甲基转移酶基因的例子（节 10.4.2；图 10.20）］。

人类中差异性启动子使用最好的例子之一涉及巨大的抗肌萎缩基因，后者总共包含 79 个以上的外显子，分布于 Xp21 内超过约 2.4 Mb 的 DNA 上。至少有七种不同的选择性启动子可供使用。三种上述选择性启动子位于通常的起始位点附近，包括一个大脑皮层特异性启动子、一个位于下游 100 kb 的肌肉特异性启动子以及一个用于小脑 Purkinje 细胞且位于下游更远 100 kb 处的启动子（图 10.14）。这些启动子的使用将产生分子质量为 427kDa 的大型异构体（称为 Dp427，这里 Dp=抗肌萎缩蛋白，并常常给一个后缀来指出组织特异性，例如 Dp427m，表示肌肉特异性异构体）。三种 Dp427 异构体区别于其 N 端最末尾的氨基酸序列，作为使用第一外显子的三种不同选择的结果。除编码通常的大型异构体的选择性启动子之外，至少还有四种其他的选择性内部启动子可供使用，产生较小的异构体（图 10.14）。

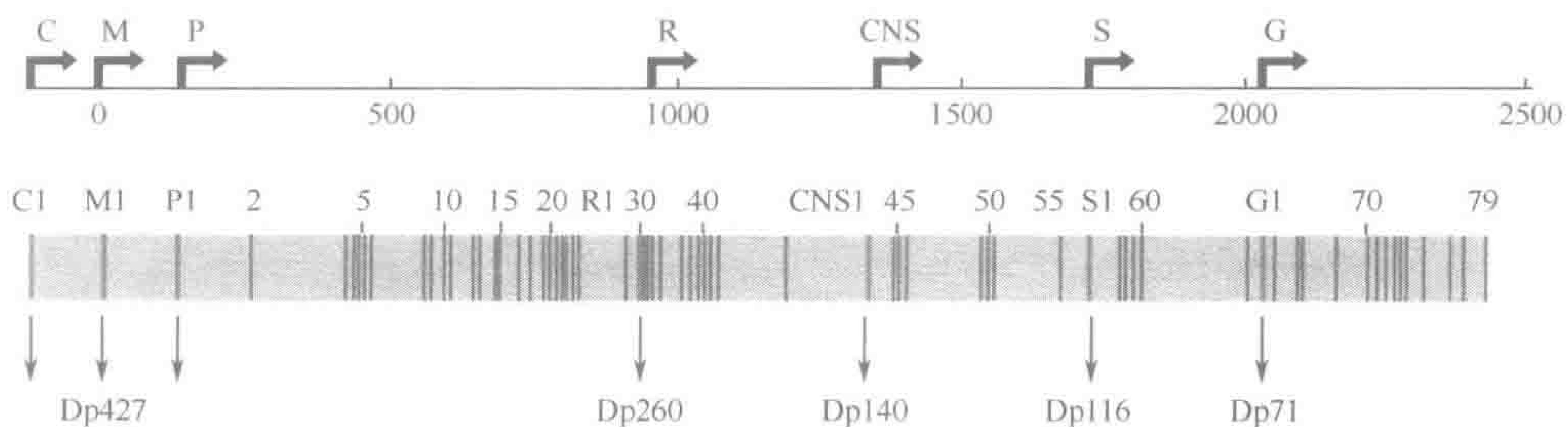


图 10.14　至少有 7 种不同的启动子可用于引起抗肌萎缩基因的组织及细胞类型特异性表达

7 种选择性启动子的位置如上方所示：C，皮质；M，肌肉；P，浦肯野细胞；R，视网膜（+脑+心肌）；CNS，中枢神经系统（+肾）；S，施旺细胞；G，常见的（几乎泛在的表达，但在完全分化的骨骼肌中检测不到）。外显子的大致位置如下方所示。**注**：每个启动子使用自己的第一外显子（红色：C1、M1、P1、R1、CNS1、S1 和 G1）加上下游的外显子（蓝色）。内部启动子位于紧邻标出的外显子的上游，具体如下：R，外显子 30；CNS，外显子 45；S，外显子 56；G，外显子 63。全长的 C、M 和 P 抗肌萎缩蛋白约为 427 kDa（Dp427）。四个内部启动子 R、CNS、S 和 G 产生依次减小的异构体：Dp260、Dp140、Dp116 和 Dp71。选择性剪接据知可发生，特别是在 3′端；更多的信息见http://www.dmd.nl/isoforms.html。

10.3.2　人类基因易于发生选择性剪接以及选择性多聚腺苷酸化

选择性启动子的使用通常涉及在转录起始处使用选择性的外显子，但其他机制，特别是选择性剪接亦有助于选择性外显子的宽泛使用。在一些相对复杂的生物（果蝇和人类分别拥有约 0.7 倍和 1.5 倍于一种简单的 1mm 长的虫子，即秀丽新小杆线虫的基因数量）中意外少的基因表明生物复杂性可能极其依赖于基因的选择性表达（Maniatis and Tasic，2002；Roberts and Smith，2002）。

选择性剪接：普遍性及模式

至少 50%（并且很可能高得多的比例）的人类基因将经历**选择性剪接**（alternative splicing），借此在 RNA 的加工过程中不同的外显子组合将表现于源自同一基因的转录物中。对很多基因而言，众多的异构体原则上可能产生于 RNA 水平，但常常并不十分

清楚可能的选择性转录物中有多少将具有生物学重要性（尽管一些显然是重要的，下面）。各种类型的选择性剪接均可发生，造成编码外显子与非编码外显子的选择性组合，以及共享某些常见序列的**外显子长度变异体**（exon length variant）（图 10.15）。编码多肽的基因有以下几种结局：

- **不同的蛋白质异构体**。这可能由编码外显子的选择性组合或者造成氨基酸差异的变异编码外显子所引起。有时候缺乏完整的功能性重要结构域或者重要定位信号的蛋白质的产生，具有重要的功能产物（框 10.3）。
- **不同的非翻译序列**。非编码外显子与变异非编码外显子的选择性组合会产生不同的5′或3′非翻译序列，有时候还会产生不同的多聚腺苷酸化位点（图 10.15）。不同的非翻译序列在功能上的意义通常不是很清楚，但一些基因中这种现象却十分显著——例如，作为选择性剪接的结果，生长激素受体 mRNA 显示至少八种不同的5′UTR 序列（Pekhletsky *et al.*,1992）。

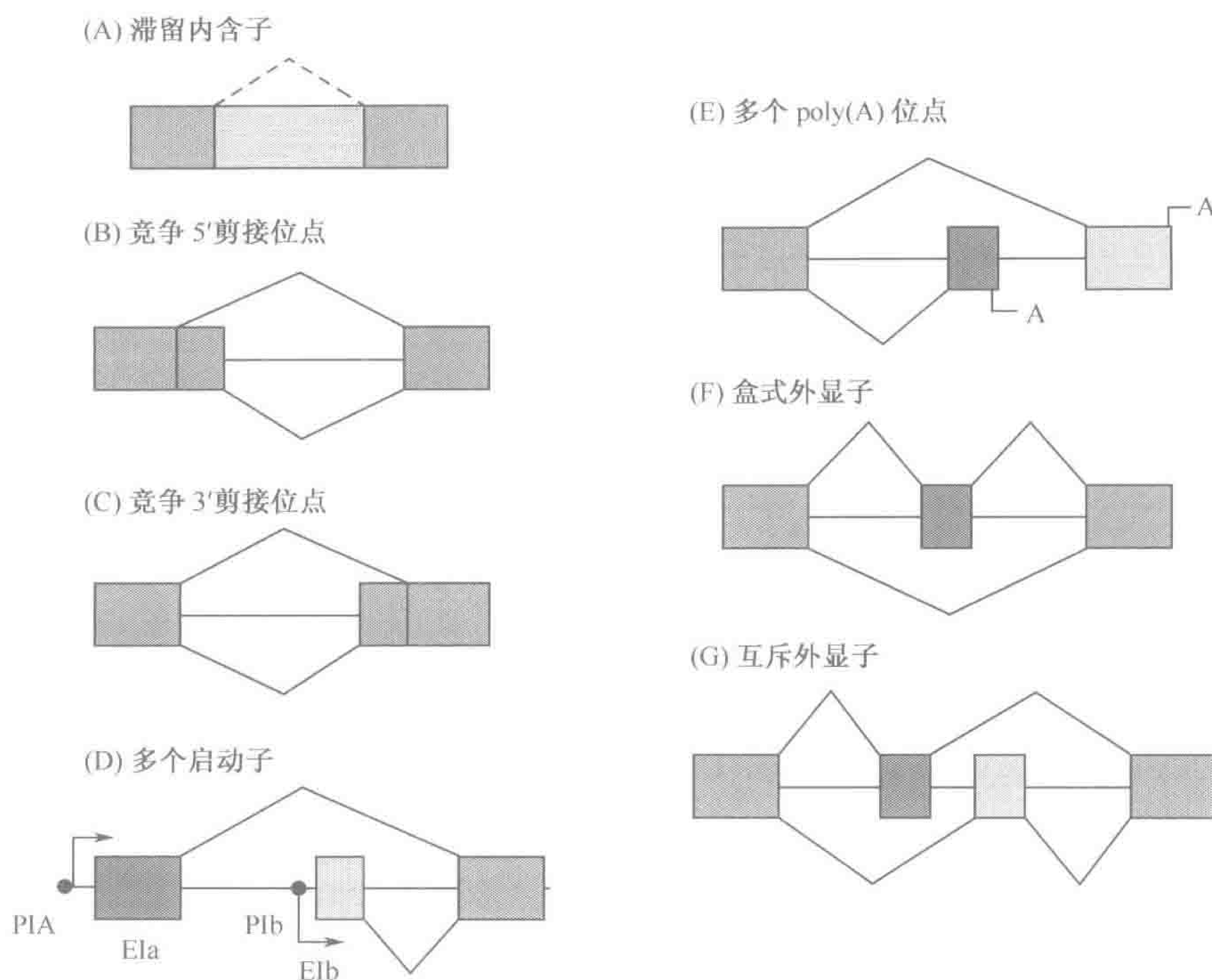


图 10.15 选择性剪接事件的类型

内部的盒式外显子（F）可独立于其他外显子而被包括或省略（跳过）[**外显子跳跃**（exon skipping）]，而**互斥外显子**（mutually exclusive exon）（G）则出现在两个或多个外显子的序列中，其中每次将仅有一个被选择包含进成熟的 RNA 中。用以说明部分这些现象的例子如下：盒式外显子，*WT1* 基因的第 5 外显子（图 10.16A）；由竞争性 5′剪接位点造成的外显子长度变异（B），*WT1* 基因第 9 外显子的变异（图 10.16A）；互斥外显子，*Dscam* 基因第 4、6、9 或 17 外显子的变异（图 10.16B）。不同启动子的选择也会引入不同的 5′外显子（图 10.14）。经 Elsevier 允许，改编自 Robert 和 Smith(2002). Curr. Opin. Chem. Biol. 6，375～383。**注**：除选择性剪接之外，在人类基因组中也发现了**反式剪接**（*trans*-splicing）的偶然例子。选择性剪接是转录自一条 DNA 链上单一转录单位内不同组合的外显子的序列结合到一起，而顺式剪接则是转录自不同 DNA 链上不同转录单位所属的外显子的序列结合到一起（Finta and Zaphiropoulos，2002；Maniatis and Tasic，2002）。

选择性剪接可产生数量巨大的潜在异构体，包括对于果蝇的 *Dscam* 基因而言多至惊人的 38 016 种蛋白质异构体（Schmucker *et al*.,2000；图 10.16B）。已证实，一些这样的选择受发育阶段和组织特异性的调节。在一些已深入研究的基因中，选择性剪接的形式已证明具有惊人的保守性，诸如 Wilms 瘤 *WT1* 基因的＋KTS 和－KTS 异构体在亲缘关系很远的生物诸如河豚鱼中也是保守的（图 10.16A；框 10.3）。

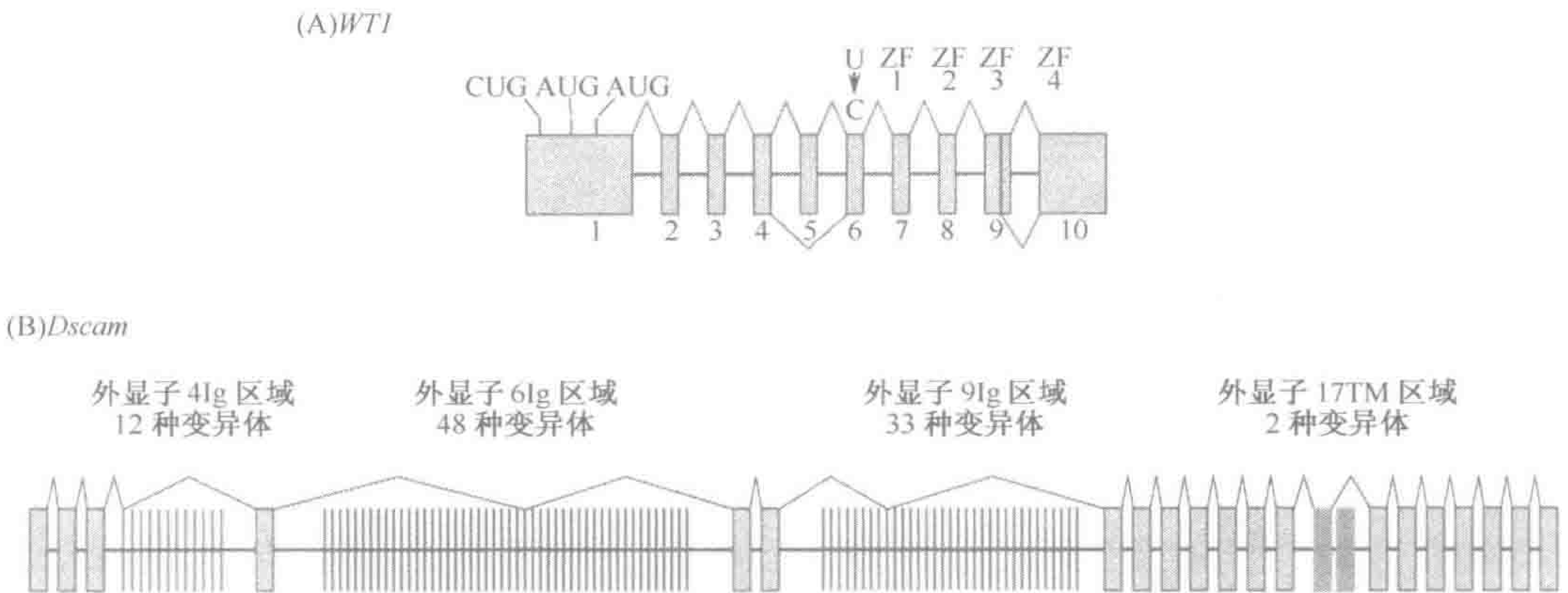


图 10.16 分别由 *WT1* Wilms 瘤基因和果蝇 *Dscam* 基因中的选择性剪接所致的功能差异和大量的潜在复杂性

（A）*WT1* 剪接。由于合并的选择性使用第 1 外显子中的三个不同起始密码子，一个第 6 外显子中的 U→C RNA 编辑替代以及两个选择性剪接活动：第 5 外显子不同的删除（跳跃）和第 9 外显子长度的变异（由第 9 内含子竞争性的 5′剪接位点所致），将可能产生 24 种异构体。第 9 外显子的变异将造成 KTS 肽的包含或删除。＋KTS 异构体被特异性地定位于细胞核内的剪接体位置，并认为在结合剪接因子中具有作用，而－KTS 异构体则更广泛地分布于核质中，并且可能在含有通用转录因子的结合结构域中具有更普遍的作用（Larsson *et al*.,1995）。（B）*Dscam* 剪接。通过选择对于编码免疫球蛋白样结构域的第 4、6 以及 9 外显子及编码跨膜结构域的第 17 外显子各自相互排斥的变异体，将能产生共 38 016（12×48×33×2）种可能的异构体（Schmucker *et al*.,2000）。

经 Elsevier Science 允许，改编自 Roberts 和 Smith(2002). Curr. Opin. Chem. Biol. 6，375～383。

框 10.3 选择性剪接能改变蛋白质的功能特性

以下所列远非全面，仅仅为了示意一个蛋白质的生物学特性能够通过选择性剪接而改变的某些途径。更多的信息见 Lopez(1998) 以及 Graveley(2002)。

- ▶ **组织特异性异构体**（tissue-specific isoform）如原肌球蛋白与降钙素的异构体（后者包括表达甲状腺的降钙素以及神经降钙素基因相关多肽）。
- ▶ **膜结合及可溶性异构体**（membrane-bound and soluble isoform）。蛋白质定位可通过产生众多膜受体的可溶形式来调控，如Ⅰ类和Ⅱ类 HLA、IgM、CD8、生长激素受体、IL-4、IL-5、IL-7、IL、促红细胞生成素、G-CSF、G-MCSF、LIF（白血病抑制因子）以及 FAS 凋亡信号受体等。
- ▶ **选择性细胞内定位**（alternative intracellular localization）。编码一种 C 端具有四个锌指结构且具有多至 24 种不同异构体的 *WT1* Wilms 瘤基因提供了一个有用的例子。差异性剪接可导致一段编码三个氨基酸、支配细胞核定位的 KTS 序列的包含或删除（图 10.16A）。
- ▶ 改变了的功能。*WT1* 基因产物的＋KTS 和－KTS 异构体在它们与靶基因中特定 DNA 序列相结合的能力方面也有所不同。前者被认为在结合剪接因子方面具有一定的作用；后者则可能在与包含通用转录因子的结构域相结合的方面具有更为广泛的作用。其他的例子包括：转录因子异构体（依靠包含或排除于蛋白质产物内的结构域的性质来激活/抑制转录；见 Lopez，1998）；以及各种基因促进凋亡和诱导凋亡的异构体，如同 Ich-1（半胱天冬酶 2）基因的情况。

选择性剪接：调节

为理解剪接调控研究得最为深入的模式系统为果蝇中的性别决定通路，后者也调节基因剂量。选择性剪接被用于这种通路的各个分支以控制转录调节子或影响转录的染色质相关蛋白的表达，剪接的正调控和负调控都很明显（Lopez，1998）。在哺乳动物细胞中候选的剪接调节子为 RNA 结合蛋白的 **SR 家族**（SR family）〔后者具有独特的富含丝氨酸（S）-精氨酸（R）二肽的 C 端结构域〕以及一些 HnRNP（异质性细胞核核蛋白颗粒）蛋白。这些蛋白据知能够促进剪接体集结过程中的不同步骤，而且它们据知亦与**剪接增强子序列**（splicing enhancer sequence）相结合，后者为可增强剪接位点识别的调节序列（Blencowe 2000；Caceres and Kornblihtt，2002；Faibrother *et al.*,2002）。

选择性多聚腺苷酸化

在人类 mRNA 中选择性多聚腺苷酸化信号的使用亦很常见，并且已发现不同类型的**选择性多聚腺苷酸化**（alternative polyadenylation）（Edwards-Gilbert *et al.*,1997）。在许多基因中，两个或更多的选择性多聚腺苷酸化信号被发现于 3′UTR 内，而选择性多聚腺苷酸化的转录物可显示组织特异性；在其他情况下，选择性多聚腺苷酸化信号可在选择性剪接后发挥作用。

10.3.3　RNA 编辑是一种罕见的加工形式，借此碱基特异性的改变被导入 RNA

RNA 编辑是一种转录后加工的形式，可涉及酶介导的 RNA 水平的核苷酸插入或缺失或者单个核苷酸的替换。插入或缺失 RNA 编辑似乎是动基体目原生动物（kinetoplastid protozoa）（诸如锥虫）与黏液菌的线粒体中基因表达的奇特性质。**替换 RNA 编辑**（substitution RNA editing）在一些系统中频繁使用，诸如在维管植物的线粒体及叶绿体中单个 mRNA 可经历多次 C→U 或 U→C 编辑事件。

在哺乳动物中并无插入或缺失 RNA 编辑的证据，但在有限数量的基因中观察到了替换编辑（Gerber and Keller，2001）。已知可发生的 RNA 编辑大多涉及非常严格的胞嘧啶或腺嘌呤残基的脱氨基（氨基基团的去除，由一个 RNA 依赖性脱氨基酶家族催化），但转氨基（通过获得一个氨基基团的修饰）亦可发生，如同在 Wilms 瘤基因中 U→C编辑的情况（图 10.16A）。已知的两类以脱氨基为基础的 RNA 编辑为：

- **C→U 编辑**。这仅发生在很少的基因中，特别是人载脂蛋白基因 *APOB*，其中对于这种编辑研究得很深入。在肝脏中 *APOB* 基因编码一个 14.1 kb 的 mRNA 转录物和一个 4536 个氨基酸的产物 apoB100。然而，在肠中一个特异性胞嘧啶脱氨基酶 APOBEC1 将第 6666 位的核苷酸由胞嘧啶转变为尿嘧啶，因此造成一个提前出现的终止密码子。截短的（7 kb)mRNA 编码一种产物 apoB48，在序列上与 apoB100 的前 2152 个氨基酸完全一致（图 10.17）。激活诱导的脱氨基酶（AID），一种涉及免疫球蛋白 DNA 重组和突变的酶与 APOBEC1 具有相当程度的相似性，却似乎用于脱氧胞嘧啶的脱氨基（节 10.6）；
- **A→I 编辑**。这种编辑由 ADAR(**a**denosine **d**eaminase **a**cting on **R**NA）脱氨基酶家族（作用于 RNA 的腺嘌呤脱氨基酶）成员催化，作用于一些编码配体闸控的离子通道

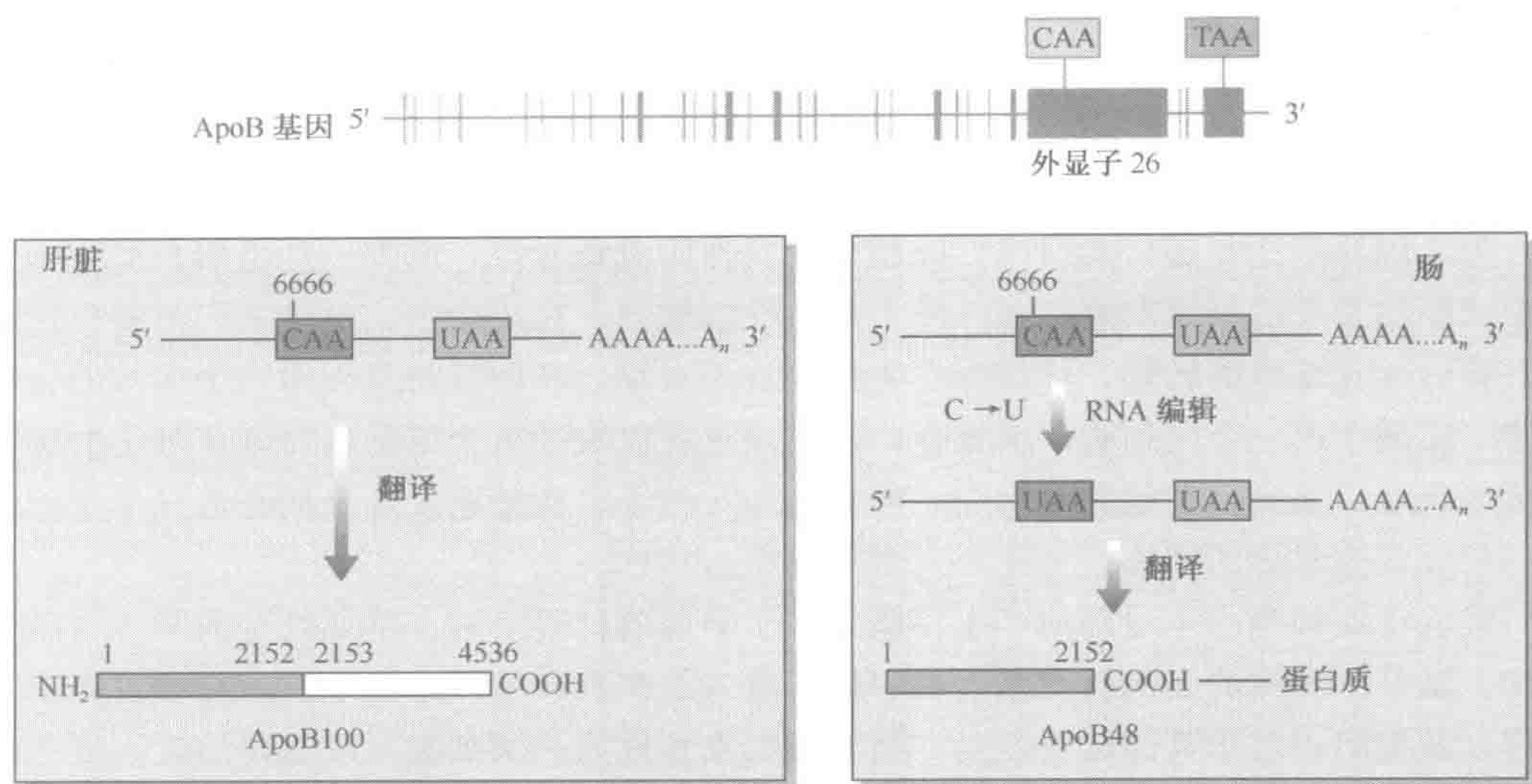


图 10.17　人载脂蛋白 B 基因（*APOB*）加工过程中的组织特异性 RNA 编辑
在肝脏中位于 APOB mRNA 6666～6668 核苷酸位置上的 2153 位密码子（CAA）编码谷氨酰胺。然而在肠中第 6666 位的 C→U RNA 编辑使 CAA 密码子为终止密码子UAA 所取代，导致一种较短的产物 ApoB48。

包括谷氨酸盐受体以及相关蛋白质的 mRNA 上。腺嘌呤脱氨基后形成**次黄嘌呤**（inosine）（I），一种通常不存在于 mRNA 中的碱基（腺嘌呤碳 6 位的氨基基团被一个 C=O 羰基基团所取代）。次黄嘌呤的行为与鸟嘌呤相似，它将优先与胞嘧啶配对，并且当出现于一个密码子中时，在蛋白质合成过程中就如同 G 一样被翻译。例如，对于谷氨酸盐受体 B 基因而言，RNA 编辑将 CAG（谷氨酰胺）密码子替换为 CIG，后者将如同 CGG 一样被翻译，产生精氨酸。这类产生 Gln→Arg 的编辑通常称为 Q/R 编辑（得名于所涉及的两个氨基酸的单字母密码）。

尽管其全部意义尚不清楚，但 RNA 编辑对于哺乳动物的生存至关重要（例如，敲除个别小鼠的 *ADAR* 基因可产生严重的表型）。或许它出现于进化过程中以纠正发生在基因组水平的错误，但它也提供了一种调节表达的形式以及产生蛋白质多样性的另外途径。

10.4　差异性基因表达：起源于不对称并由诸如 DNA 甲基化等表观遗传机制得以永存

人类基因表达的组织特异性的概念建立已久。知之甚少的是这种模式最初是如何建立起来的。由于一个有机体中所有的有核细胞内的 DNA 组成实际上完全一致，遗传机制将无法解释差异性的基因表达最初是怎样在细胞中出现的。为了解释这个现象，Waddington 援引了发育过程中基因调控的**表观遗传机制**（epigenetic mechanism）。遗传机制将解释由 DNA 序列改变（突变）引起的可遗传状态（特征），而表观遗传机制则描述不依赖于 DNA 序列的可遗传状态。近来，多种表观遗传机制被发现在脊椎动物

的细胞中起作用，包括一些可使体细胞谱系中基因表达的特定状态得以永续的机制。

10.4.1 哺乳动物胚胎细胞中基因的选择性表达很可能是对近程的细胞-细胞信号事件的反应

为了解释表达在随后的组织、细胞以及发育阶段特异性，需要一些机制在受精卵细胞中或胚胎发育的极早期建立起一种不对称性或轴向。在果蝇中，由于基因产物自位置不对称的抚育细胞的转移，卵细胞在先天上就不对称。胚胎最初是作为一个多核的合体细胞（实质上的一个大细胞）来发育，而区域化将取决于单个细胞核对调节分子的远程梯度的反应。然而，在哺乳动物中，卵细胞相对较小，而胚胎发育早期将造成单个细胞的一种外观为对称的聚集。然而，发育将变得不对称。

哺乳动物细胞内不对称的产生可能来源于早期的位置暗示。早期发育的某些方面在先天上就是不对称的，包括受精过程中精子进入的位点、植入过程中胚胎在子宫壁上的附着以及细胞相对于其邻居的位置。随着胚胎发育成为一团细胞，以及稍后更为复杂的结构将形成，个体细胞在可用的邻近细胞的数量方面将发生变化。直接的细胞-细胞信号或近程的细胞间信号事件可提供一种确定细胞定位的手段，并引起差异性的基因表达。例如，如果某种细胞内信号分子具有比方为一个细胞直径的作用范围，那么囊胚（节 3.7.2；图 3.13）外的细胞将会收到来自各个方向的邻近细胞的不同信号，而不同的定位提示可能被转化为差异性的基因表达。由于特定的细胞系统形成于例如器官发生阶段（大多于胚胎期第 4～9 周完成），特定细胞类型的生长或分化因子可能随后诱导发育阶段及/或组织特异性转录因子的表达。

10.4.2 DNA 甲基化是脊椎动物细胞中永存的基因抑制的一种重要表观遗传因素

差异性基因表达模式一旦建立，表观遗传机制即可确保它们在细胞分裂时稳定地遗传，从而提供一种在细胞谱系中传递的**细胞记忆**（cell memory）形式。表观遗传机制可确保一些靶基因或基因组区域的转录激活状态（‘开放’染色质构象）的稳定遗传，或者是能够组织某些基因组区域的染色质采用一种高度凝缩、转录失活的形式。目前认为至少有两种，并很可能有三种不同的表观遗传机制在动物发育中起作用：

- **DNA 甲基化**（DNA methylation）——一种使染色质形成关闭的、转录失活状态的表观遗传机制（见下文）；
- **聚梳-三胸**（polycomb-trithorax）基因调节。聚梳类抑制因子与三胸类激活因子通过改变染色质的结构，形成一种‘关闭’（转录抑制）或者‘开放’（转录激活）的构象，来维持若干关键性发育调控因子（包括同源异型基因）的正确表达（Mahmoudi and Verrijzer，2001）。
- **组蛋白修饰**（histone modification）——第三种可能的表观遗传机制，可能为造成特定基因组位置永存化表达状态的原因（Turner，2002；节 10.2.1）。

DNA 甲基化是目前公认的一种重要的表观遗传机制，它与组蛋白修饰（节 10.4.3）相互作用以允许抑制基因表达的染色质状态从二倍体细胞向子细胞的传递（Bird，2002）。然而，DNA 甲基化在真核细胞中的精确功能了解得还不十分完善，并且显示出明显的种属差异（框 9.3）。脊椎动物胞嘧啶甲基转移酶将识别一种 CpG 靶序

列，但与细菌甲基化酶不同，它们对于识别半甲基化的 DNA 靶（仅有一条链已被甲基化）显示出强烈的偏好。序列 CpG 呈现二分体对称性，因此在 DNA 复制之后，新合成的 DNA 链将获得与亲代 DNA 相同的 CpG 甲基化模式（图 10.18）。因此，CpG 甲基化模式能够稳定地传递至子细胞。这种对预先存在的甲基化模式的永存化有时称为**维持甲基化**（maintenance methylation），并且在哺乳动物细胞中由 *Dnmt1* 甲基转移酶来实现。

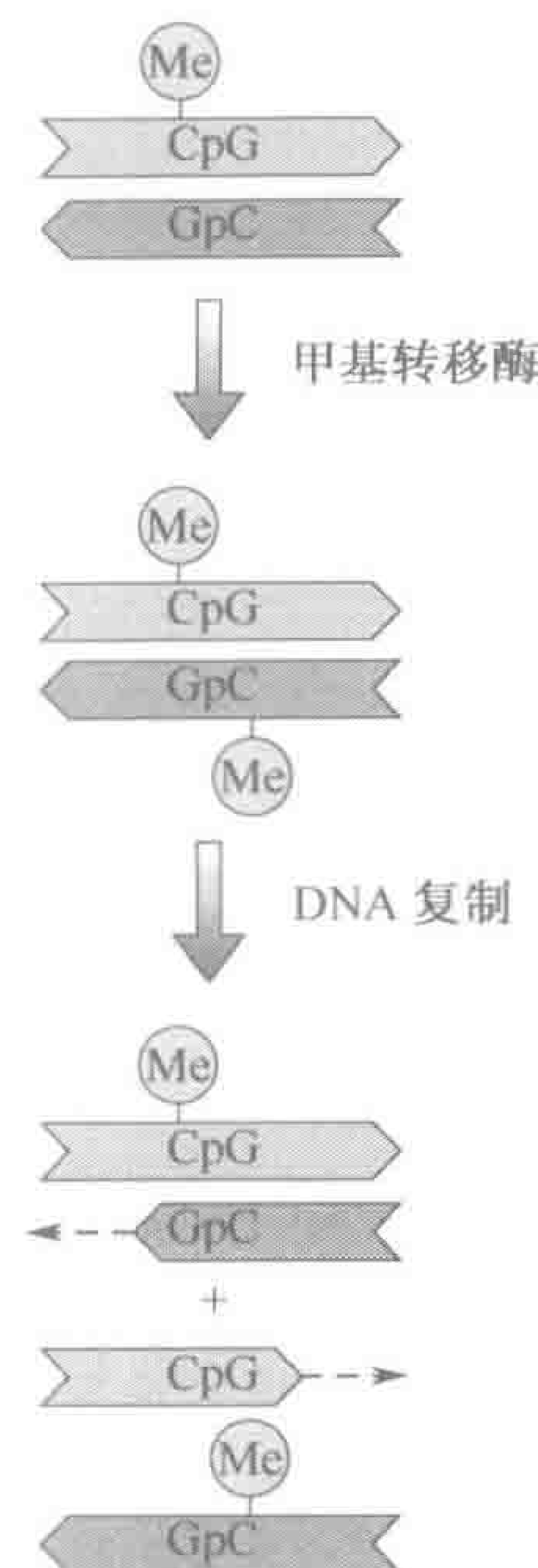


图 10.18　CpG 甲基化通过对特定甲基转移酶识别一个半甲基化的靶序列的必要条件而永存化
序列 CpG 具有二分体对称性。一个**半甲基化**（hemi-methylated）的靶序列（仅一条链被甲基化）甲基化之后，两条甲基化的链将在 DNA 复制时分离并作为合成两条非甲基化子链的模板。所产生的子代双链这时将为延续同样的甲基化模式提供新的半甲基化靶。

分化的体细胞基因组中 5-甲基胞嘧啶的分布模式因细胞类型而异，但维持甲基化将确保在单个体细胞谱系中甲基化的模式甚为稳定。然而，在早期发育中，甲基化将有显著的改变，构成某种形式的**表观遗传重编程**（epigenetic reprogramming）（Razin and Kafri，1994；Reik *et al.*，2001；Li，2002）。在发育过程中存在两种主要类型的表观遗传重编程（图 10.19）。

生殖细胞中的重编程

胚胎的原始生殖细胞（配子最终将出自这些细胞）由高度甲基化的 DNA 出发，但之后在发育过程中将发生逐渐的去甲基化。至原始生殖细胞已经进入性腺时，去甲基化已基本完成并将很快结束。然而，在性腺分化之后，当生殖细胞开始发育时，**重新甲基化**（*de novo* methylation）将发生。这将导致哺乳动物精子和卵细胞 DNA 的大量甲基化。精子基因组比卵子基因组的甲基化程度更高，而在甲基化模式上的性别特异性差异很明显，特别是在印记的基因座方面（Merteneit *et al.*，1998）。

胚胎早期的重编程

受精的卵母细胞的基因组为精子和卵细胞基因组的聚集，因此它与极早期胚胎均被充分地甲基化，并在许多基因的父源与母源性等位基因上存在甲基化的差异。稍后，在植入前胚胎的桑椹胚及早期囊胚阶段，将发生**基因组范围的去甲基化**（genome-wide demethylation）。再随后，在原肠胚形成前期，广泛的**重新甲基化**（*de novo* methylation）将实现。然而，这种甲基化的范围将因不同的细胞谱系而异：**体细胞谱系**（somatic cell lineage）为高度甲基化；**滋养层衍生的谱系**（trophoblast-derived lineage）（产生胎盘、卵黄囊等）为低甲基化；**早期原始生殖细胞**（early primordial germ cell）被省却；

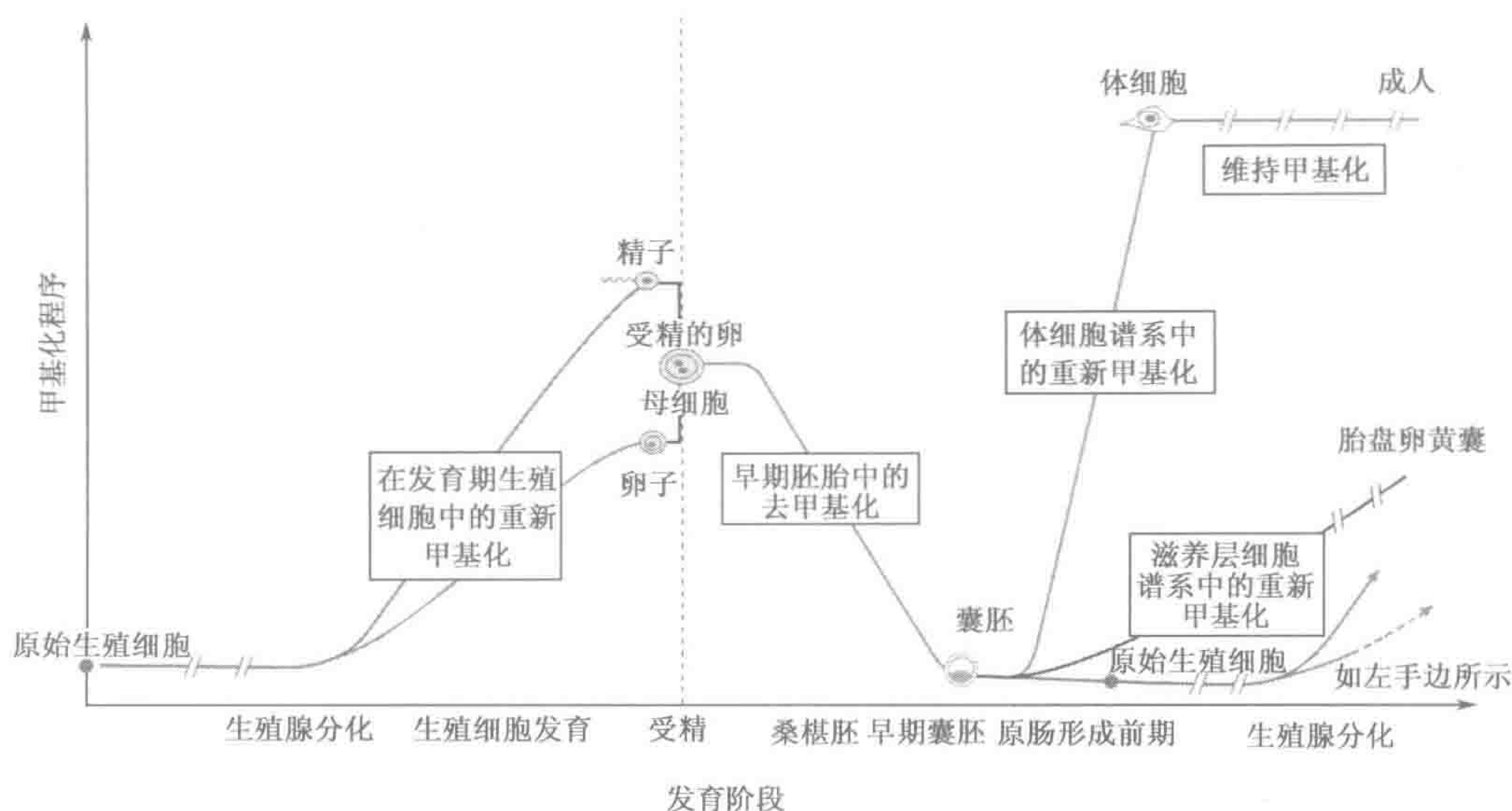


图 10.19 哺乳动物发育过程中 DNA 甲基化的改变

配子发生和早期胚胎发育的发育阶段被展开来阐述；后期发育的部分被压缩，如双斜杠所示。**注意** DNA 甲基化非常迅速地改变于：（Ⅰ）**配子发生**（gametogenesis）过程——重新甲基化将在精子和卵子中产生相当范围内甲基化的基因组（尽管在这些基因组中的整体甲基化水平以及甲基化模式上存在差异——见正文）；以及在（Ⅱ）**早期胚胎**（early embryo）中，其中一波基因组范围的去甲基化将发生于植入前阶段（桑椹胚和早期囊胚），并立刻接替以始于原肠形成前期的大范围重新甲基化。后者在体细胞谱系中尤其显著，而在形成胎盘和卵黄囊的滋养层细胞谱系则程度较轻，但不发生在原始生殖细胞（最终将形成精子和卵子细胞的胚胎细胞）中。

它们的基因组 DNA 将保持很大程度上的未甲基化直至性腺分化之后（如上所述）。

不同的甲基转移酶对于重新甲基化的作用仍不清楚。Dnmt3a 和 Dnmt3b 甲基转移酶是有力的候选者，但靠其自身仍不充分。相反，它们似乎是与 Dnmt1 甲基转移酶相互作用。*Dnmt1* 基因高度表达于男性生殖细胞、成熟的卵母细胞以及早期胚胎中，而作为使用卵母细胞特异性及精母细胞特异性启动子的结果，表达将受到性别特异性的调节（图 10.20）。

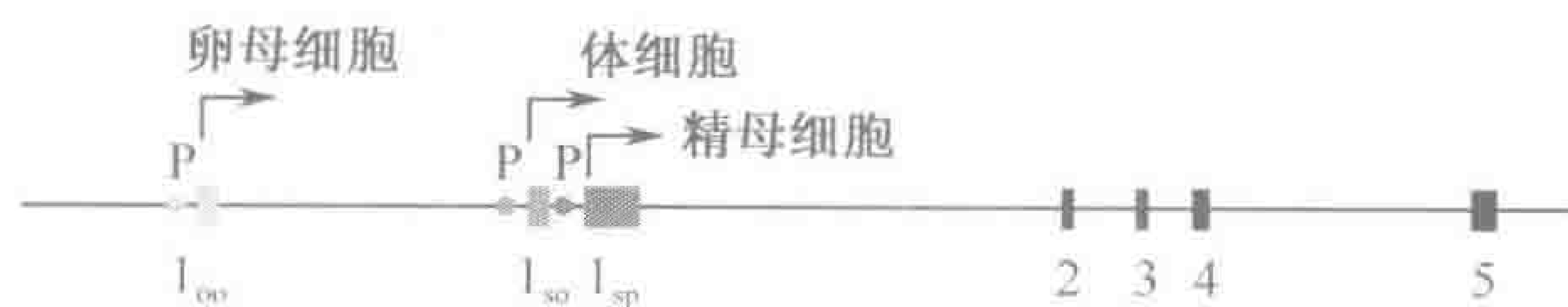


图 10.20 性别特异性启动子调节 *Dnmt1* 甲基转移酶基因

Dnmt1 甲基转移酶基因似乎为小鼠细胞中主要的维护 DNA 甲基化的甲基转移酶并且可能也是主要的重新甲基化的甲基转移酶。它高度表达于男性生殖细胞、成熟卵母细胞以及早期胚胎中。存在五个外显子，但由于选择性启动子的使用，对于第一外显子存在三种不同的可能选择（很像不同的 p427 抗肌萎缩蛋白异构体对于选择性启动子的选择，图 10.14）。它们是：**外显子** 1 **so**（somatic，用于体细胞中）；**外显子** 1 **o**（oocyte，用于卵母细胞中）；以及**外显子** 1 **sp**（spermatocyte，用于精母细胞中）。卵母细胞特异性外显子与活性 Dmnt1 甲基转移酶的大量产生相关，后者在 N 端被截短并且在生长的稍晚阶段隐遁于细胞质中。精母细胞特异性启动子将干扰翻译并且在男性减数分裂的交叉阶段中阻止 Dmnt1 的产生。（Mertineit *et al*.,1998）。

10.4.3　动物的 DNA 甲基化可能提供对转座子的防卫以及调节基因表达

尽管并非所有的真核细胞看起来都受 DNA 甲基化调控，但它在动物细胞内的功能却的确显得至关重要，而靶向敲除小鼠的胞嘧啶甲基转移酶基因将导致胚胎期的死亡。然而，动物细胞内 DNA 甲基化的确切功能仍不清楚。目前的观察特别地聚焦于动物细胞的两个方面：基因组的大小（动物细胞具有相对较大的基因组和大量的基因，并且还有大量属于转座子类型的高度重复 DNA 家族）；以及发育的模式（尤其是在寿命和细胞更新速度方面的差异）。对于动物细胞中 DNA 甲基化的主要功能，两种截然相反的观点成为大量争议的对象：**宿主防御模型**（host defense model）和**基因调节模型**（gene regulation model）。

宿主防御作为 DNA 甲基化的一种主要功能

和细菌中 DNA 甲基化的限制-修饰功能（框 5.2）一样，宿主防御模型推测动物细胞中 DNA 甲基化的主要功能是赋予一种形式的基因组保护，但在这里是抑制转座子的传播（Yoder *et al.*,1997）。人类基因组中约 45%的 DNA 序列可被归类转座子家族，而在人类基因组和其他基因组中一小部分这类序列据知在活跃地转座（节 9.5）。在人类和其他基因组中的转座子家族据知处于高度甲基化状态（约 90%的 5-甲基胞嘧啶认为是位于反转座子家族中），因此 DNA 甲基化被看作是抑制这类转座的一种机制，后者如果未被抑制，预期将会对细胞具有伤害性。然而，最近从一种无脊椎脊索动物——玻璃海鞘（*Ciona intestinalis*）所获得的数据似乎与这种基因组防御模型不符：一种外观活跃的反转座子的多个拷贝以及一大部分高度重复的 SINE 基本上未被甲基化，相对而言基因则似乎已被甲基化（Simmen *et al.*,1999）。

基因调节作为 DNA 甲基化的主要功能

脊椎动物中的 DNA 甲基化被认为是使转录沉默的一种机制，并可能构成一个默认的位置。活跃转录的 DNA 序列需要被去甲基化（至少在启动子区域）。虽然在无脊椎动物中 DNA 甲基化可能用于抑制转座子以及其他的重复序列家族，在脊椎动物中它可能已经获得了一种特殊的角色，即作为调节内源性基因的表达以及减少转录噪音（通过使一大部分在细胞中不需要其活性的基因沉默）的一种机制。

相反的意见是组织特异性基因 5′区域的甲基化状态无法与不同组织中的表达找到关联，并且甲基化在基因表达中的作用是在于使用等位基因特异性基因表达的机制（例如印记等）所产生的特化的生物学功能上（Walsh and Bestor，1999）。

DNA 甲基化与基因表达

转录活跃与不活跃的染色质 DNA 在若干特征上有所不同，包括压缩的程度及其甲基化的范围（表 10.6）。尽管启动子下游的 CpG 岛的甲基化并不阻止持续的转录通过这些区域（Jones，1999），甲基化的启动子区域却无疑与转录沉默相关。另外，**组蛋白乙酰化**（histone acetylation）的程度也是一个重要因素（另见节 10.2.1）。特定的组蛋白乙酰基转移酶将乙酰基团添加到靠近组蛋白 N 端的赖氨酸残基上，导致更为开放的

染色质构象；组蛋白去乙酰化将增进基因表达的抑制。

表 10.6　与转录活跃及不活跃的染色质相关的特征

特征	转录活跃的染色质	转录不活跃的染色质
染色质结构	开放、延伸的构象	高度浓缩的构象；在异染色质中尤其明显（包括兼性异染色质和组成性异染色质）
DNA 甲基化	相对地未甲基化，尤其在启动子区域	甲基化，包括在启动子区域
组蛋白乙酰化	乙酰化的组蛋白	去乙酰化的组蛋白

DNA 甲基化与组蛋白修饰的过程之间存在联系（Li，2002）。启动子区域中甲基化 CpG 序列处的抑制似乎是由特异性地结合至甲基化 CpG 的蛋白所介导。已发现两种这类蛋白，MeCP1 和 MeCP2（**me**thylated **C**pG-binding **p**rotein，甲基化 CpG 结合蛋白 1 和 2），而后者已证实为胚胎发育所必需，并作为转录抑制因子发挥作用。MeCP2 部分通过恢复 HDAC 的活性来沉默基因的表达，导致染色质重构。乙酰基基团自 H3K9（组蛋白 3，第 9 位的赖氨酸残基）被除去，接着在 MeCP2 辅助下 H3K9 甲基化，组蛋白甲基化聚集诸如导致染色质浓缩的蛋白质如 HP1 的信号（Fuks *et al*.，2002 以及其中的参考文献和图 10.21）。

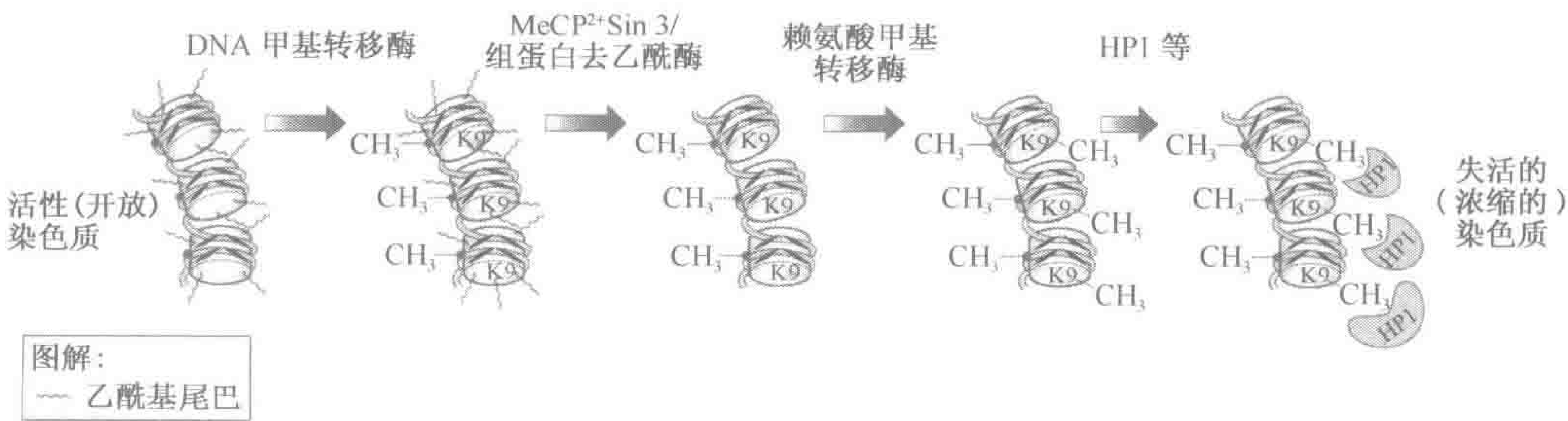


图 10.21　DNA 甲基化可通过组蛋白去乙酰化和组蛋白 H3K9 甲基化介导转录抑制

CpG 双核苷酸为 DNA 甲基化的目标，而甲基化的 CpG 又是蛋白质如 MeCP2 特异性结合的目标。MeCP2 充当一种转录抑制因子，并可聚集一种由转录因子抑制物 mSin3A 及组蛋白脱乙酰酶组成的辅抑制物复合体。MeCP2 也可结合组蛋白甲基转移酶，因此当 H3K9（组蛋白 3 第 9 位上的赖氨酸）去乙酰化后，它将甲基化。甲基化的 H3K9 是异染色质化的蛋白质诸如能够引起染色质浓缩和转录失活的 HP1 的作用目标（Fuks *et al*.，2003）。

10.5　基因表达的远程控制与印记

10.5.1　染色质结构可对基因表达施加远程控制

与细菌基因不同，真核基因通常是个别转录的。启动子和相关的上游元件通常控制一个基因的表达，拥有距这些调控元件 1 kb 以内的转录起始点。然而，一些顺式作用元件将跨越大得多的染色体区域来施加远程控制，并且协同调控基因簇的证据越来越多。

对于基因被重新定位至基因组中其他地方的观察亦提示或许是由于组蛋白修饰（节10.2.1），染色体组织成为基因表达的功能性结构域（**染色质结构域**，chromatin domain）。例如，当基因被转移到新的染色体区域时（或者是由于染色体的自发断裂事件，或者是由于转基因实验），异常的基因表达可能经常发生，即使整个基因与紧邻序列中必需的控制序列被完整地保留了下来。相邻的染色体结构域被推测由**绝缘子**（insulator）（又称**边界元件**，boundary element）充当远侧增强子和沉默子作用的屏障（Bell *et al*.,2001）。

对于增强子或沉默子的竞争

基因表达的远程控制有时似乎依赖于成簇的基因对一个增强子的竞争。这似乎是珠蛋白基因表达的一个特征，如节10.5.2所述。

异染色质诱导的位置效应

果蝇中染色体重排的研究显示与着丝粒、端粒或异染色质板块靠近可抑制基因表达，这大概是通过改变大的染色质结构域的结构（异染色质诱导的位置效应）所致。类似的位置效应在人类中尚未得到详细地刻画，但跨越大的染色体结构域控制基因表达的远程效应的证据已通过对人类中疾病相关染色体断裂点的研究而呈现。具体的例子为导致无虹膜症和弯肢性发育不良，但却距受损的疾病位点*PAX6*和*SOX9*数百kb的染色体断裂（Kleinjan and van Heyningen，1998）。

Prader-Willi以及Angelman综合征（框16.6）集中了位置效应、印记和DNA甲基化。一个类似于珠蛋白基因座控制区的顺式作用序列已被发现，后者将支配15q11上的亲代特异性甲基化以及百万碱基大小的染色体区域的基因表达。

X失活

哺乳动物中X染色体失活似乎是由一个基因*XIST*发起的，后者独特地表达于失活的X染色体之上（节10.5.6）。这种作用尚未被理解，但肯定是由某种远程的染色质结构改变所介导。这是因为一种可扩散的*XIST*介导的因子将不能只影响*XIST*基因表达所在的X染色体。

10.5.2 基因簇中的个体基因的表达可由共同的基因座控制区域来协调

一些人类基因簇呈现簇中的个体基因协调表达的证据。例如，α珠蛋白、β珠蛋白以及4个*HOX*基因簇中的个体基因将按对应于它们在染色体上的线性队列的时间顺序依次被激活。对珠蛋白基因而言，阶段特异性的表达与发育过程中血红蛋白产物的更迭的位置相对应。因此，在胚胎发育早期，血红蛋白将产生于胚外膜，即卵黄囊，但在让位给成人中的骨髓之前，在胎儿中肝脏将成为主要的合成部位。这种发育进展为两个主要的珠蛋白基因簇各自中基因的开启与关闭所伴随，产生形式略有不同的血红蛋白（**血红蛋白转换**，hemoglobin switching；图10.22）。

珠蛋白基因簇（以及其他一些基因簇）中基因的表达已被认为由位于基因簇上游一段距离的**基因座调控区**（locus control region，LCR）来协调。LCR是通过它们在转基

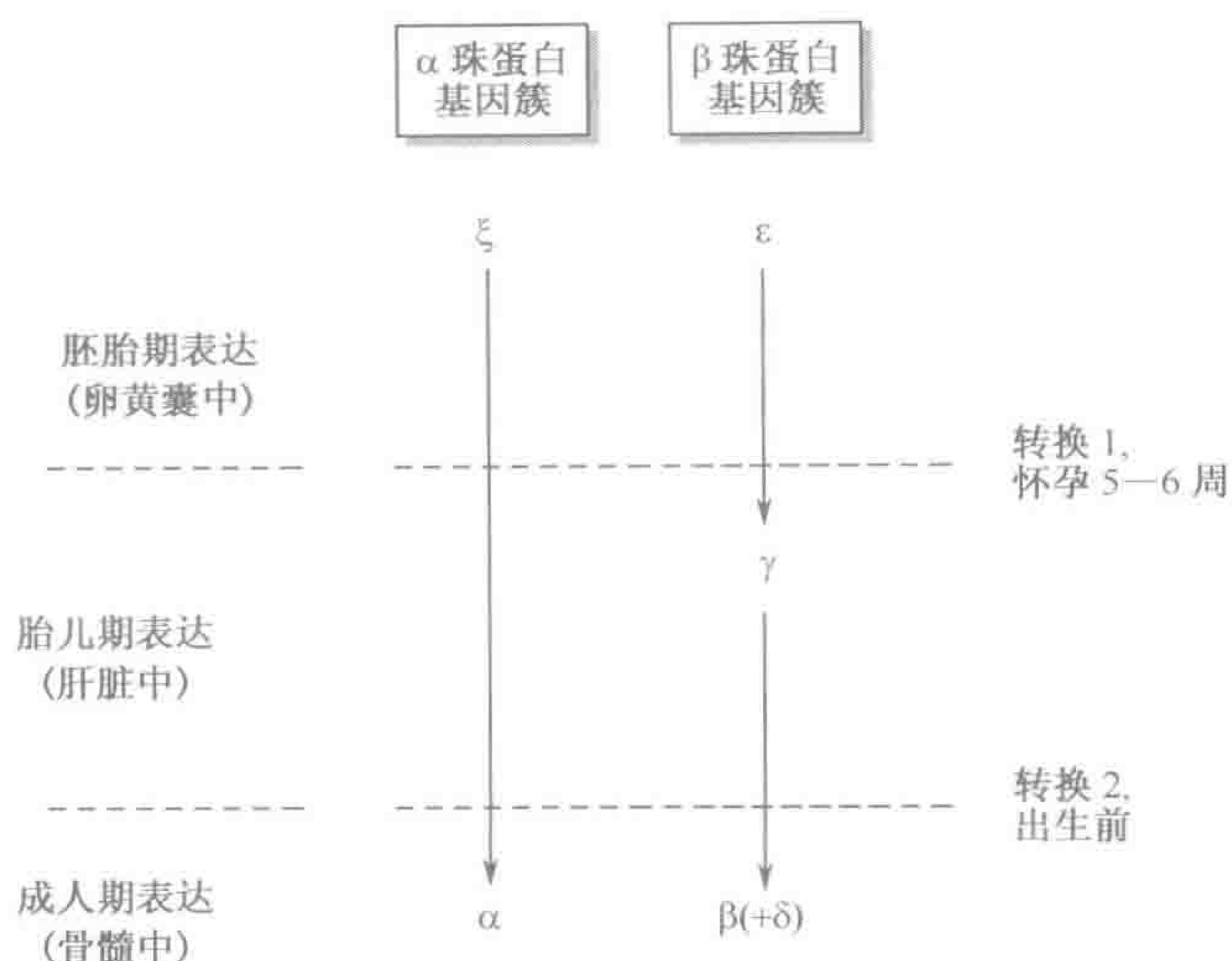


图 10.22 人类血红蛋白转换发生于两个截然不同的发育阶段

因试验中的能力而明确的显性调控区：LCR 与检测基因或基因簇相连接然后转染，并让其整合至另一个基因组中。当这些发生时，LCR 能够指引相连基因在所有调查过的整合位点上高表达（并且以每基因拷贝适度恒定的水平）。相比之下，当增强子与基因相连，并用同类方式研究时，相连基因将因整合位置而呈现非常可变的表达，因而增强子将受到负位置效应的影响。因此，LCR 由于能够激活转录而与增强子相似，但它们亦由于能够在整合位点压倒负调控信号而具有一种显性效应。

由一个异位的（与其在基因组中的正常位置不同的任何位置）LCR 驱动的高水平表达似乎是由于两种可分的功能："开放的"活性染色质结构域的建立和直接的基因激活。具有转录活性的染色质结构域的开放构象使它们更易于被 DNA 酶 Ⅰ 的切割所接近。与这种关系相一致，人类 β 珠蛋白的 LCR 被认为由发现于红细胞系统 DNA 中的五个主要的 **DNA 酶 Ⅰ 高敏感位点**（DNase Ⅰ-hypersensitive site）上的短序列组成（但未见于来自不明显表达珠蛋白基因的细胞的 DNA 中）。DNA 酶 Ⅰ 高敏感位点成簇分布于 β 珠蛋白基因上游 50～60 kb 处的一个 10 kb 区域内，而 α 珠蛋白的 LCR 则被定位于一个红细胞系统特异性 DNA 酶高敏感位点 HS-40，位于 α 珠蛋白基因上游 60 kb 处（图 10.23A）。

由于含有许多泛在及组织特异性转录因子的结合位点，这类 DNA 酶 Ⅰ 高敏感位点上的 DNA 序列与增强子序列类似（图 10.7）。

其他的 DNA 酶 Ⅰ 高敏感位点位于珠蛋白基因的启动子上，但呈现发育阶段特异性。例如，在胎儿肝脏中，两个 γ 基因、即 β 和 δ 基因的启动子以 DNA 酶 Ⅰ 高敏感位点为特征，但是在成人骨髓中，这两个 γ 基因将不再具有转录活性，而它们的启动子也不再展现 DNA 酶 Ⅰ 高敏感位点。珠蛋白基因表达的发育阶段特异性转换因此被认为是通过珠蛋白基因之间在与它们各自的 LCR 相互作用方面的竞争以及基因特异性沉默元件的阶段特异性激活而实现的。例如，ε 珠蛋白基因（*HBE1*）的转录在胚胎阶段将优先受到邻近 LCR 的刺激。然而，在胎儿中，ε 珠蛋白基因的表达将随着一个沉默子的

激活而被抑制，而 γ 珠蛋白基因的表达将成为主流（图 10.23B）。

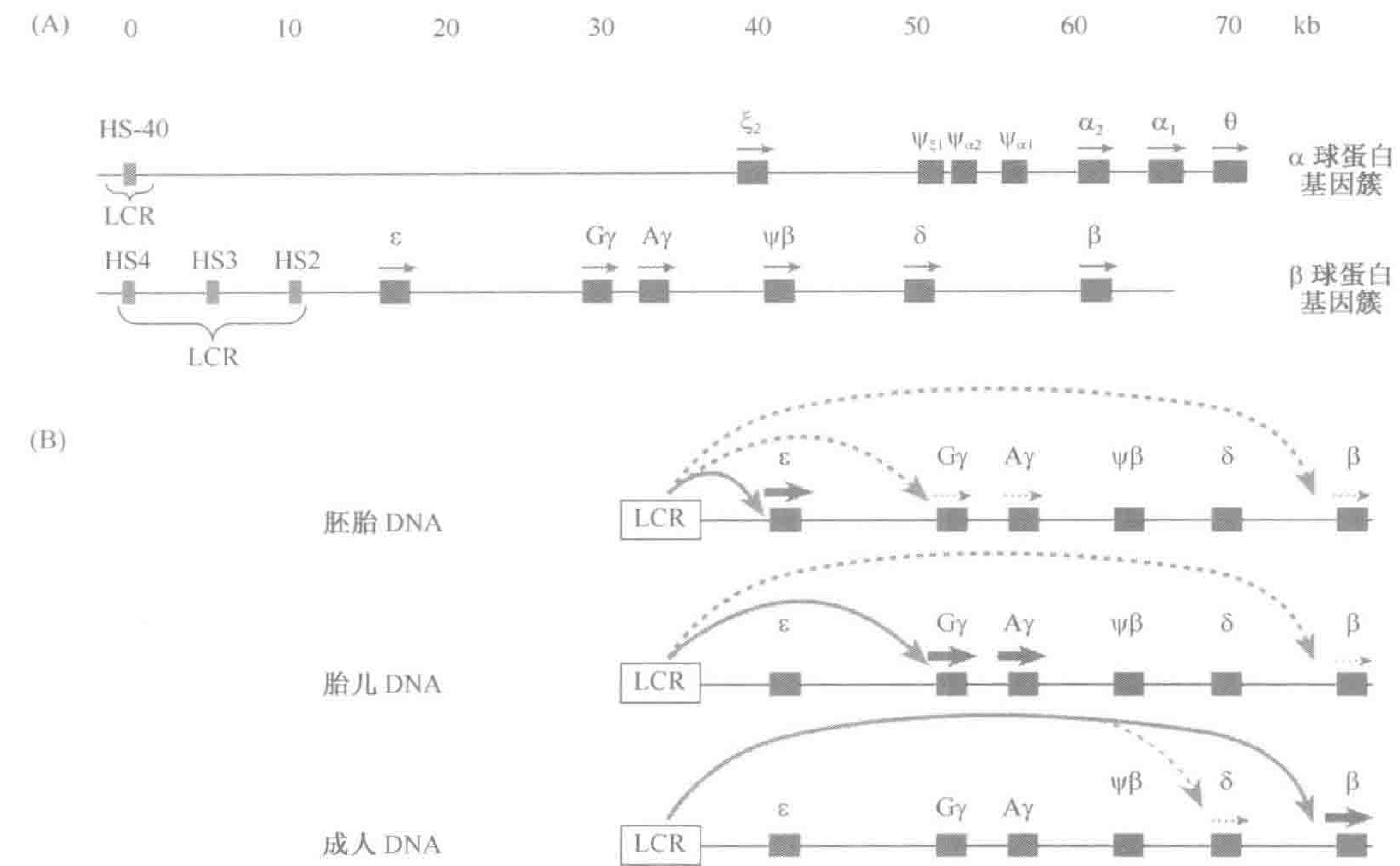


图 10.23　α 和 β 珠蛋白基因簇中的基因表达可被共同的基因座调控区控制

（A）**人类 α 珠蛋白和 β 珠蛋白基因簇的结构**。基因座调控区（LCR）由位于基因簇上游的一个或多个红细胞系统特异性 DNA 酶 Ⅰ 高敏感位点（HS-40 等）组成。箭头标明了表达基因的转录方向。θ 珠蛋白基因的功能状态尚不确定：它被表达，但 θ 珠蛋白却未被掺入到任何血红蛋白分子中。（B）**提出的 β 珠蛋白 LCR 对基因表达的调控**。实线箭头表示由 LCR 作用于所示基因上的强大增强子效应，导致高表达水平；虚线箭头表示相对应的弱效应。对于该模型尚存在一些争议——见正文。

倘若没有各自的 LCR，珠蛋白基因的表达将可忽略不计，但上述调控机制的准确性质仍然不甚清楚。早期的模型推测圈除间隔的 DNA 使 LCR 元件与下游基因的特定启动子发生直接的实体接触以激活这些基因，但简单的 LCR-启动子相互作用现在看起来不太可能（Bulger and Groudine，1999）。更让人惊讶的是，虽然人类 β 珠蛋白的 LCR 能够诱导染色质在异常位置上松弛，但当它在 β 珠蛋白基因簇中原本的位置上时并不具有这样的功能，在那里它只能像一个简单的转录增强子那样起作用（Bulger *et al.*,2002）。这些，加上最近对于显著的物种差异的了解，意味着这一领域是不断发展的，建议有兴趣的读者参阅最新的文献。

10.5.3　一些人类基因选择性地仅表达两个亲代等位基因中的一个

女性中的 X 连锁基因以及所有的常染色体基因均为双等位基因，因为父亲和母亲在正常情况下将各贡献一个等位基因。在拥有一条 X 染色体和一条 Y 染色体的男性中，绝大多数性连锁基因为单等位基因性：X 染色体上许多的基因中的大多数在 Y 染色体上没有功能性同源体；而 Y 染色体上很少的基因中的一些据知为 Y 特异性，诸如 *SRY*，即主要的男性性别决定基因座（然而，也已知在少数情况下功能性基因同源体发

现于X和Y染色体上，节12.2.8)。

我们习惯于设想双等位基因的父源性和母源性基因均被表达，除非其中一个或两个拷贝均遭受了影响表达的突变。表达无疑可能为组织和细胞类型特异性，但通常在两个等位基因的功能性能方面并没有本质差异。然而，在人类和其他哺乳动物中，已知有若干双等位基因，其中一个亲代等位基因，或父源性或母源性，但不是两个，正常情况下在一些细胞中被抑制（构成了一种形式的**等位基因排斥**，allelic exclusion)。在这类细胞中相关基因被认为展现**功能性半合子状态**（functional hemizygosity)：正常情况下仅获得最大基因产量的一半，即使两条亲代等位基因的序列与正常基因表达完全相符乃至于完全一致。在一些情况下，等位基因排斥可能是特选细胞或组织的特性，而在同一个体的其他细胞中，两个等位基因则可能均正常表达。

尽管最初被认为是一种稀有事件，双等位基因的**单等位性表达**（monoallelic expression）已被证明于越来越多的人类基因中。多种不同的表达机制可能与之有关，而两大类机理涉及其中（Chess，1998；Ohlsson *et al*.,1998)；

- **取决于亲代来源的等位基因排斥（印记）**。在一些情况下，两个遗传的拷贝中选择哪一个表达并非是随机的。这就意味着对一些基因而言，表达抑制的等位基因总是父源遗传的等位基因；在另一些中则总是母源遗传的等位基因（节10.5.4)；
- **不依赖于亲代来源的等位基因排斥**。在这里，两个等位基因中哪一个被抑制最初是随机的，但后来等位基因排斥的模式将随着细胞分裂稳定地传递到子细胞中去。可能涉及多种不同的机制（框10.4)。

框10.4　导致人类细胞中双等位基因的单等位性表达的机制

机制	相关基因的例子以及单等位性表达的细胞定位
A. 取决于亲代来源的等位基因排斥	
基因组印记	少数基因（印记基因目录的细节见进一步阅读)。细胞定位将取决于单个基因表达的位置，但是注意某些印记基因在一些细胞类型中呈单等位性表达，但在另一些中则呈双等位性表达（表10.7)。
B. 随机的等位基因排斥（不依赖于亲代来源)	
由X失活引起的等位基因排斥	仅限于女性中特定的X连锁基因。自未失活的X染色体的等位性表达仅存在于表达该基因的细胞中（节10.5.6)
在程序性DNA重排之后的等位基因排斥	B淋巴细胞中免疫球蛋白基因的表达；T淋巴细胞中T细胞受体基因的表达（节10.6.3)
由不明机制引起的等位基因排斥	神经元中的嗅觉受体基因；NK细胞受体基因；特定的白介素基因（*IL2*，*IL4*）*XIST*（在早期女性胚胎的细胞中)；*PAX5*（在成熟的B细胞和早期的始祖细胞中)

10.5.4　基因组印记涉及取决于亲代来源的等位性表达的差异

对于哺乳动物的多方面观察已表明在一个个体中母源与父源基因组并不等价（框

10.5)。除了精子基因组与卵母细胞基因组 DNA 的遗传差异外，还存在表观遗传的差异。一个主要的差异在于 DNA 甲基化的总量（精子基因组比卵母细胞基因组具有更为广泛的甲基化）以及特定 DNA 序列种类中 DNA 甲基化的模式。例如，LINE1 序列在精子细胞中被高度甲基化，但在卵母细胞中却仅被部分甲基化（Razin and Kafri，1994；Yoder *et al*.,1997)。同样，在一些单个的基因座上，父源与母源等位基因的甲基化程度存在重大差异。例如，父源性的 *H19* 等位基因高度甲基化，而母源性的等位基因却存在甲基化不足。

框 10.5　母源与父源基因组的不等价性

除了明显的 X/Y 染色体的差异外，父源与母源遗传的常染色体及 X 染色体之间的不等价性显示于下列的观察上。

实验诱导的小鼠中的单亲二倍体

受精的小鼠卵母细胞中的雄性原核可被去除并被另一个雌性原核所取代，从而产生一个**雌源体**（gynogenote）[有时称为**孤雌体**（parthenogenote)；所有 38 条染色体均为母源性]。相反，倘若雌性原核被另一个雄性原核所取代，则将形成一个雄源体。尽管拥有正常的二倍体染色体，这类胚胎将无法发育并在孕中期以前死亡。雌源体将呈现胚胎外结构的严重缺陷，但相对正常的胚胎；与此相反，在雄源体中，胚胎将比胚胎外结构受到更为严重的影响（Bestor，1998)。

人类中自然发生的单亲二倍体（另见节 2.5.4）

人类的单亲孕体并非罕见。雄源孕体将发育为**葡萄胎**（hydatidiform mole)，后者由成团的水肿绒毛及其他胎盘结构组成，但缺乏胚胎组织。雌源孕体将形成皮样囊肿，后者将发育成为**卵巢畸胎瘤**（ovarian teratoma)，由一团分化良好但高度杂乱的成体组织构成，常包括骨骼、牙齿、软骨、皮肤以及其他组织，但通常没有任何胚胎外结构。

三倍体流产儿可视为从一个亲代获得一个二倍体基因组并从另一个亲代继承一个正常的单倍体基因组的组合。其表型将因哪个亲代提供了二倍体的基因组而有所不同。

单亲二体性（另见节 2.5.4）

一些孕体具有正常的 46，XX 或 46，XY 核型，但可能从两个亲代中仅一个那里获得了同一条染色体的两个拷贝。这可能导致因相关染色体的亲代来源而异的异常表型。例如，从他们的父亲那里继承了 15 号染色体的两个拷贝的 46，XX 或 46，XY 个体将发生 Angelman 综合征；如果两个拷贝均为母源遗传，则将导致 Prader-Willi 综合征（框 16.6)。

亚染色体突变根据亲代来源而产生差异性的异常表型

- 当位于母源或父源染色体上时，特定染色体区域的缺失将产生不同的表型。最好的例子是 15q12 的缺失，后者在父源染色体上将产生 Prader-Willi 综合征，而在母源染色体上则将产生 Angelman 综合征（框 16.6)。
- 某些人类性状为常染色体显性遗传，但仅显露于继承自一个亲代时。在一些家族中血管瘤以常染色体显性的方式遗传，但仅表现于从他们的父亲那里继承了这个基因的人中。Beckwith-Wiedemann 综合征（MIM 130650）有时为显性遗传，但仅表现于那从他们的母亲那里继承了它的人中。家系的例子如图 4.5D、4.5E 所示。
- 许多肿瘤中的等位基因丢失（第 17 章）将优先涉及父源等位基因。

正如框 10.5 中所表示，父源与母源基因组之间的差异将导致父源与母源等位基因在表达上的差异。哺乳动物中的**基因组印记**（genomic imprinting）（又称配子印记

或亲代印记）描述了特定的基因座上取决于亲代来源的等位性表达的不等价性的情况（Reik *et al.*,2001；Sleutels and Barlow，2002）。在所有（或至少一些）表达这些基因的组织中，或者是父源遗传的等位基因，或者是母源遗传的等位基因的表达将始终被抑制，导致单等位性表达。相同模式的单等位性表达可随细胞分裂而准确无误地传递给子细胞。然而，由于表达受到抑制的等位基因的核苷酸序列有可能与基因表达完全相符（并且甚至可能与表达的等位基因完全一致），因此这是一种表观遗传现象，而不是遗传现象。

印记的普遍性与进化

大多数人类基因并不受印记的支配，否则我们将不会见到这么多简单的孟德尔性状。已开展系统的调查以发现小鼠中印记的染色体区域。与人类不同，所有的小鼠染色体均为近端着丝粒型，而罗伯逊异位将容许形成交叉，产生某一特定染色体的两个拷贝均来自一个亲代的子代（**单亲二体**，uniparental disomy，UPD；节 2.5.4）。这些揭示了一些染色体的 UPD 并不具有表型效应；而另一些则将产生异常表型。有时这些异常表型对于不同的亲代来源是互补的，例如，生长过度常见于母源性 UPD，而生长迟缓则见于父源性 UPD。对某些染色体而言，UPD 将为致死性。

在染色体和基因水平的进一步剖析显示印记是个别基因或小的染色体区域的一种特性。目前，已知在人类和小鼠中共有大约 60 种印记基因，包括编码多肽的基因以及编码功能性非编码 RNA 的基因（Tycko and Morrison，2002）。这些基因已知的生理功能可能呈现相当大的差异，但提示性的证据表明其中许多基因将调控生长以及神经行为学性状。在人类基因组中已知有两个主要的印记基因簇：11p15.5（包含 Beckwith-Wiedemann 综合征区域）上的一个含有至少八个印记基因的 1 Mb 区域（Marher and Reik，2000）；以及一个位于 15q11-q13 区域（包含 Prader-Willi 以及 Angelman 综合征区域）、含有十个以上印记基因的 2.2 Mb 基因簇（Meguro *et al.*,2001；图 10.24）。

绝大多数已知的印记基因位于常染色体上。然而，在建立 X 染色体失活（见下节）中发挥主要作用的 *XIST* 基因可能被看作是印记的 X 连锁基因的一个例子，因为母源遗传的等位基因的表达将优先在滋养层细胞中被抑制。一种影响认知功能的的印记 X 连锁基因亦认为由于 Turner 综合征中不同的行为模式。患 Turner 综合征的女孩没有 Y 染色体，但仅有一条 X 染色体。如果这条 X 染色体继承自母亲，社交方面常见有破坏性行为，但如果继承自父亲，这些女孩则将表现更接近于正常的行为（Skuse *et al.*，1997）。

印记据知可发生在种子植物、一些昆虫以及哺乳动物中。照表型来判断，并未在一些模式生物诸如果蝇、秀丽小杆线虫以及斑马鱼中观察到主要的印记效应，尽管印记的潜势可能在果蝇中存在。哺乳动物因胚胎完全依赖于来自母体胎盘的丰富营养而非同寻常。由于许多印记基因均参与调节胎儿的生长，一种解释设想了**亲代基因组冲突**（parental genome conflict）：父源基因组通过产生一个积极地从母亲那里移走营养的胚胎来最好的传播它自己；而母源性基因组则将抑制这种活动以保护母亲，并为将来的后代节省一些资源。如同在单亲二倍体（框 10.5）中之所见，父源性基因将优先表达于滋养层细胞及胚外膜中，而母源性基因则优先于表达于胚胎中。

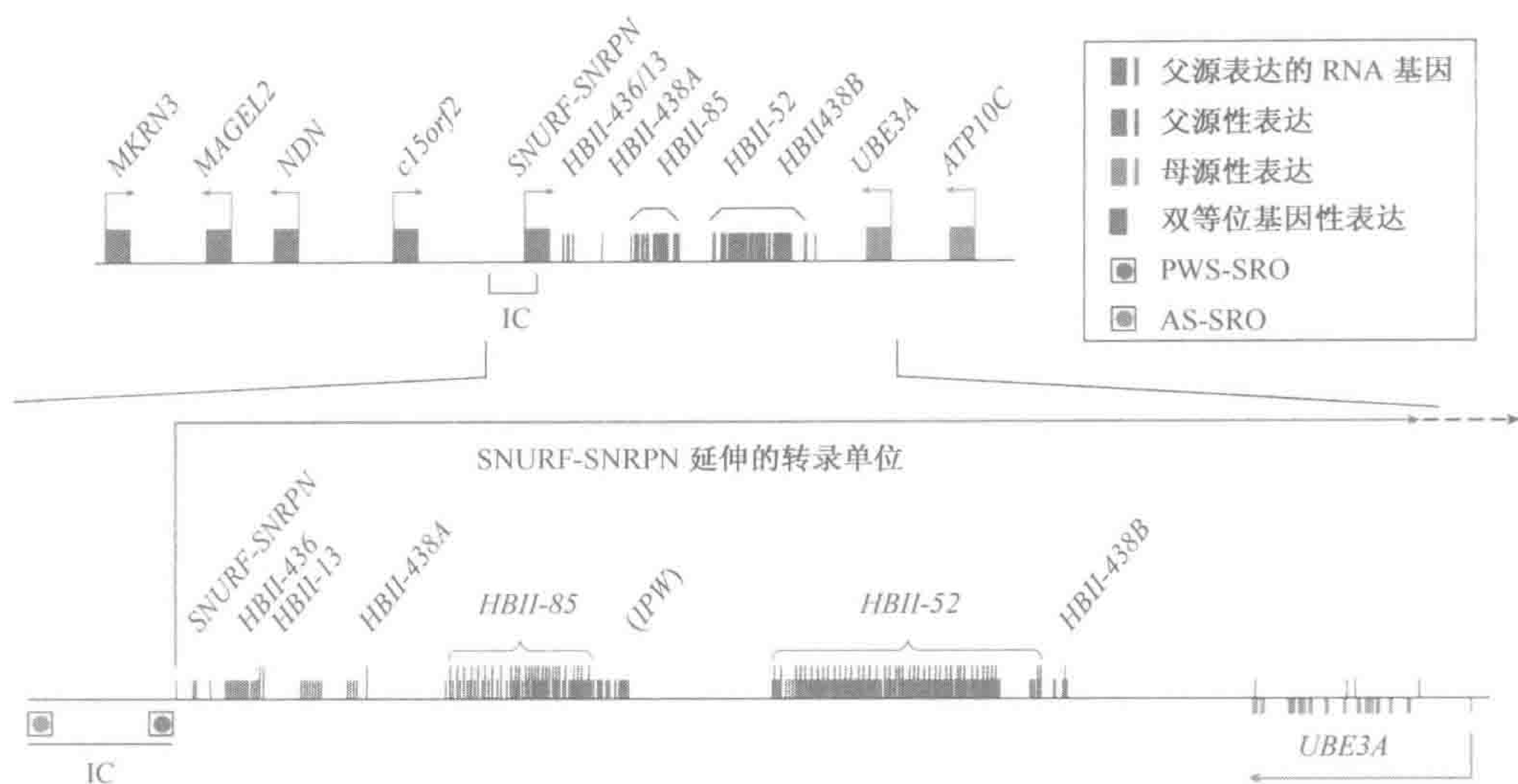


图 10.24　在 15q11-q13 上的 Prader-Willi 综合征/Angelman 综合征相关印记基因簇

箭头表示转录方向。印记的编码多肽基因包括 *UBE3A* 和 *ATP10C*，二者优先表达自母源性 15 号染色体，而 *MKRN3*、*NDN* 以及 *MAGEL2* 则优先表达自父源性的 15 号染色体。另外，复杂的 *SNURF-SNRPN* 转录单位（>148 个外显子；延伸超过 460 kb，与 *UBE3A* 基因以及可能还有 *ATP10C* 基因重叠）也被印记。它编码两种蛋白质（由外显子 1～3 编码的 SNURF 蛋白以及由外显子 4～10 编码的 SNRPN 剪接子蛋白）外加一些 RNA 转录物（此处未显示，但在文献中被称作 *IPW*、*PAR5*、*UBE3AS* 等），并可能作为一种反义 RNA 调节子来调控 *UBE3A* 和 *ATP10C* 基因。除了这些复杂性以外，共有 79 个 snoRNA 序列位于 *SNURF/SNRPN* 转录单位的内含子中，包括各一份拷贝的 *HBⅡ-436*、*HBⅡ-13*、*HBⅡ-437*（可能是个假基因）、两份相同但分开的 *HBⅡ-438* 拷贝，27 份 *HBⅡ-85* 拷贝以及 47 份 *HBⅡ-52* 拷贝（用大的竖线条表示，相对于表示 *SNURF/SNRPN* 外显子的小竖线条，Runte *et al*.,2001）。几乎所有这些拷贝（*HBⅡ-437* 除外）均为父源性表达，尤其是在脑中，并且它们可能在 Prader-Willi 综合征中具有直接的作用（Gallagher *et al*.,2002）。IC，imprinting center。经牛津大学出版社允许，改编自 Runte 等（2001）. Hum. Mol. Genet. 10（23）：2687～2700。

10.5.5　基因组印记的机制尚不清楚，但一种关键的组分似乎为 DNA 甲基化

为了证实一个基因的印记，有必要确定一名对某种出现于成熟 mRNA 中的序列变异为杂合性的个体；然后可在来源于不同组织的 mRNA 中检查单等位性或双等位性表达，并通过验明亲代的类型来确定每个等位基因的来源。对于一些基因而言，这种分析已显示印记仅局限于特定的组织或特定的发育阶段（表 10.7）。因此，印记将允许从额外的水平对基因表达进行调控，但并无可能将它的机能压缩成为简单而统一的情节。

表 10.7　哺乳动物中印记基因的组织及发育阶段调控的例子

基因	受抑制的等位基因	表达模式的差异
IGF2（胰岛素样生长因子 2）	母源性	在许多组织中印记，但双等位性表达于大脑、成人肝脏、软骨细胞等中
PEG1/MEST	母源性	在胎儿组织中印记，但双等位性表达于成人血液中
UBE3A（泛素蛋白连接酶 3）	父源性	仅在大脑中印记；双等位性表达于其他组织中
KvLQT1（钾通道）	父源性	在几种组织中印记，但双等位性表达于心脏中
WT1（Wilms 瘤基因）	父源性	在胎盘和大脑的细胞中印记，但双等位性表达于肾脏中

印记基因通常发现组织成为基因簇，包藏有**印记调控元件**（imprint control element），一类跨越长距离作用的顺式作用调节元件。在一个这类的基因簇中通常可发现一些呈偏向于父源性表达的基因紧邻于另一些呈偏向于母源性表达的基因（例子见图10.24）。主要的印记调控元件已发现局限于称为**印记中心**（imprinting center）的较小DNA区域。对位于15q11-q13的Prader-Willi综合征（PWS）/Angelman综合征（AS）而言，一个印记中心已限定在*SNURF-SNRPN*基因的5′端的上游并延伸至其中。它含有两个印记调控元件：在*SNURF-SNRPN*基因启动子和5′端的PWS-SRO元件，后者将负责建立并维持父源性印记；以及位于*SNURF-SNRPN*上游约135 kb的AS-SRO元件，后者为造成母源性印记的原因（Perk *et al*.,2002）。

印记基因簇通常亦含有编码非翻译RNA的基因，后者的表达常常与邻近的多肽编码基因的抑制相关连。PWS-AS基因簇提供了许多例子，但已知还有诸如11p15.5内的*H19*基因等的其他例子。尽管临近印记的多肽编码基因的印记RNA基因的功能在大体上尚不清楚，但被预期为一种调控作用，而直接的证据可得自至少一个例子中：小鼠的*Air*基因已被证实可调控印记的*Igf2r*（胰岛素生长因子Ⅱ受体基因；Rougeulle and Heard，2002；Sleutels *et al*.,2002）。

以上观察表明某些机制肯定能够区分父源与母源遗传的基因：当染色体经男性和女性种系传递时，它们必须获得一些印记以在正在发育的生物体中显示父源和母源等位基因的区别。一个关键的组分，至少在维持印记状态方面，就是等位基因特异性的DNA甲基化：所有印记基因均以差异性甲基化的富含CG的区域为特征，而若干基因的印记化已证明在其有*Dnmt1*胞嘧啶甲基转移酶——一个主要的维持甲基化的甲基化酶的基因缺陷的突变体小鼠中被破坏。

有意思的是，*Dnmt1*据知具有性别特异性的外显子（节10.4.2；图10.20）。在卵母细胞中这将导致一种卵母细胞特异性的N端截断的蛋白质产物，后者可以想像将能够特异性地使诸如胰岛素样生长因子Ⅱ受体之类基因的母源性等位基因甲基化。*Dnmt1*的精母细胞特异性外显子将干扰*Dnmt1* mRNA的翻译，而父源特异性的甲基化模式是如何获得的仍不甚清楚。在发育过程中，上述印记预期应该可以在至少多轮的DNA复制中稳定地遗传（例外情况见下文）。很明显，在种系发育中必然也存在在需要的时候用于消除这种印记的某种机制，例如，当一个男子将继承自他的母亲的一个等位基因传递下去时（图10.25）。发生于早期胚胎中、造成原始生殖细胞实质上未甲基化的去甲基化作用（图10.19）是能够实现这一点的一种途径。

10.5.6 哺乳动物中的X染色体失活涉及对基因表达很长范围的顺式作用抑制

X染色体失活的本质

X染色体失活（X chromosome inactivation）为发生于所有哺乳动物中的一个过程，导致在雌性中两条X染色体之一上的等位基因的选择性失活（Lyon，1999）。它提供了一种**剂量补偿**（dosage compensation）机制，后者将克服常染色体（A）基因剂量与X染色体基因剂量在预期比值上的性别差异。携带一条X染色体的雄性将仅具有X连锁基因的一个等位基因，因而是X连锁基因的组成性半合子。因此在A∶X基因剂量的

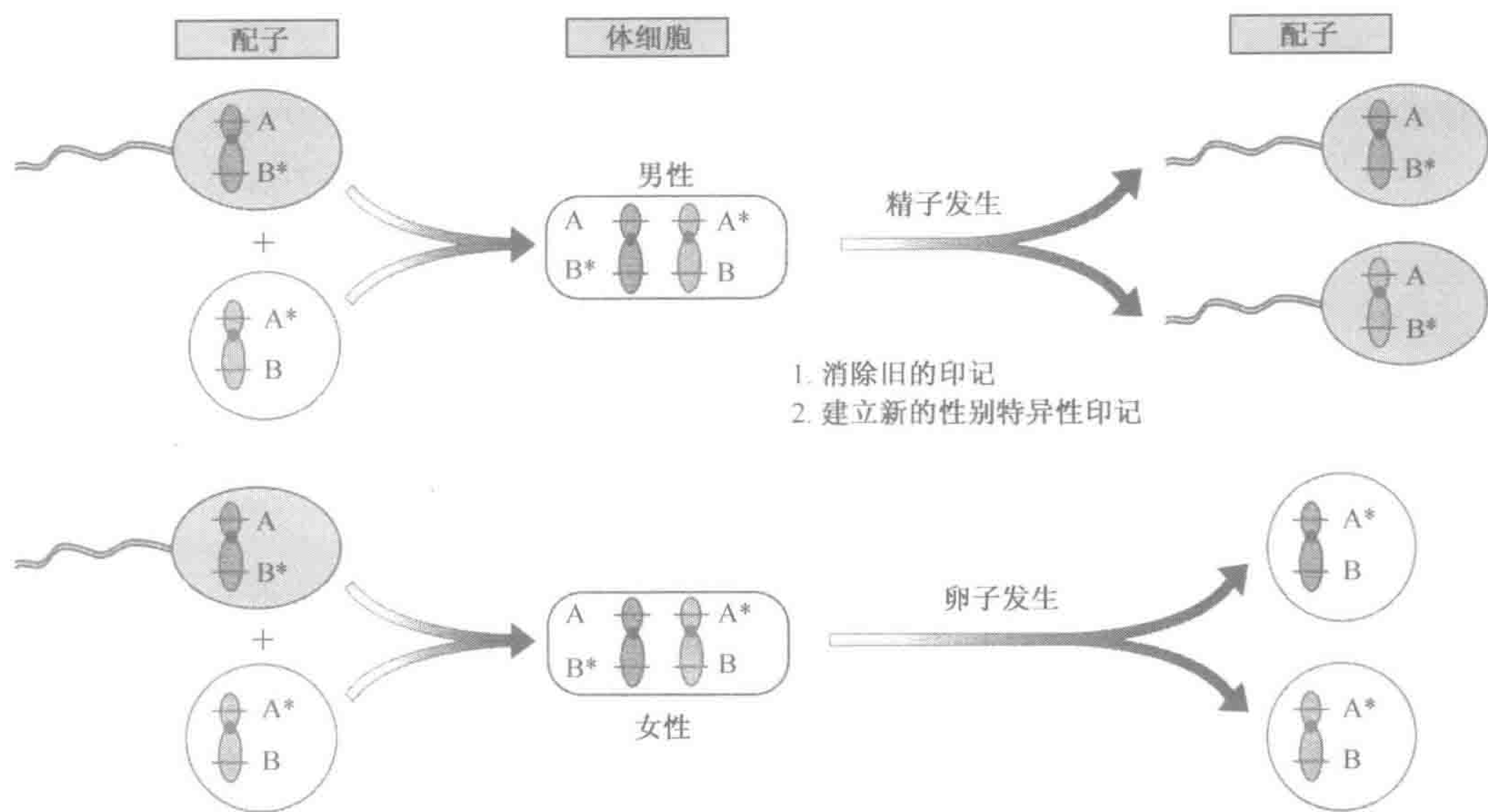


图 10.25　基因组（配子）印记需要消除种系中的印记

本图示意了携带经受印记的两个基因 A 和 B 的染色体的命运：A 将在女性种系中被印记，B 在男性种系中印记，如星号所示。因此，在二倍体细胞中 A 将在位于母源遗传的染色体上时印记，而 B 则将在出现于父源遗传的染色体上时印记。单条染色体可能经连续世代中男性及女性种系传递：一个男子可能传递一条继承自其母亲的染色体，而一个女子则可能传递一条继承自其父亲的染色体，如右框中的配子所示。因此，必然存在一种机制，借此在建立新的性别特异性印记之前将旧的印记从种系中消除。

比值上存在性别差异（雄性为 2∶1，而雌性则为 1∶1）。剂量比非常重要，因为在多种重要的代谢及发育途径中，X 连锁基因的产物需要与常染色体基因的产物相互作用，并且对关键的**剂量敏感基因**（dosage-sensitive gene）而言，对产物的数量存在严密的调控。

为了抵消 A∶X 基因剂量的这种性别差异，在大多数雌性哺乳动物的细胞中将作出一种补偿调节：两条亲代 X 染色体之一将失活。X 染色体失活涉及染色质结构的修饰，导致一种浓缩的异染色质化的结构，即 Barr 小体（可见于雌性细胞的核膜内沿）。失活的 X 染色体上的大多数基因将受到某种机制的作用，使它们发生转录失活。因此，通过使两条亲代 X 染色体之一失活，雌性哺乳动物将成为对大多数 X 连锁基因而言的**功能性半合子**（functionally hemizygous）。并不是在失活的 X 染色体之上的所有基因均失活；逃脱了 X 失活的极少数基因包括那些在 Y 染色体上存在功能性同源体，以及一些基因剂量似乎并不重要的基因（逃脱了 X 失活的基因的例子见节 12.2.8）。

X 染色体失活的时机选择

在发育的极早阶段，两条 X 染色体均有活性，但是当细胞开始由全能或多潜能细胞分化时，即于小鼠、并且很可能也在人类囊胚阶段的后期，X 失活即被启动。在每个将要形成雌性胎儿的细胞中，两条亲代 X 染色体之一将被选择以失活，但失活父源继承的 X 染色体（X^p）抑或母源继承的 X 染色体（X^m）的选择通常将会是随机的，因此也会因细胞而异［注意：滋养层细胞和有袋类哺乳动物则不同：X^p 染色体将优先失活，

即一个组织局限性印记的例子，在携带 X：常染色体易位的个体中，结果将是正常的 X 染色体一致失活]。X 染色体失活的模式在世代传递的过程中需要被消除，或许是作为这种机制的一部分，X 染色体据知在雄性和雌性的配子发生过程中均短暂地失活。

一旦早期胚胎中的始祖细胞已专注于失活 X^p 或 X^m 染色体，这种失活模式将呈**克隆式遗传**（clone inheritance）：所产生的细胞谱系中的所有后裔细胞均具有与始祖细胞一样的 X 失活模式（图 10.26A）。这就意味着所有雌性哺乳动物均为嵌合体，含有失活父源性 X 染色体细胞系和失活母源性 X 染色体细胞系的混合体。这将为三色猫的嵌合性皮毛颜色模式的照片所例证（图 10.26B）。

X 染色体失活的机制

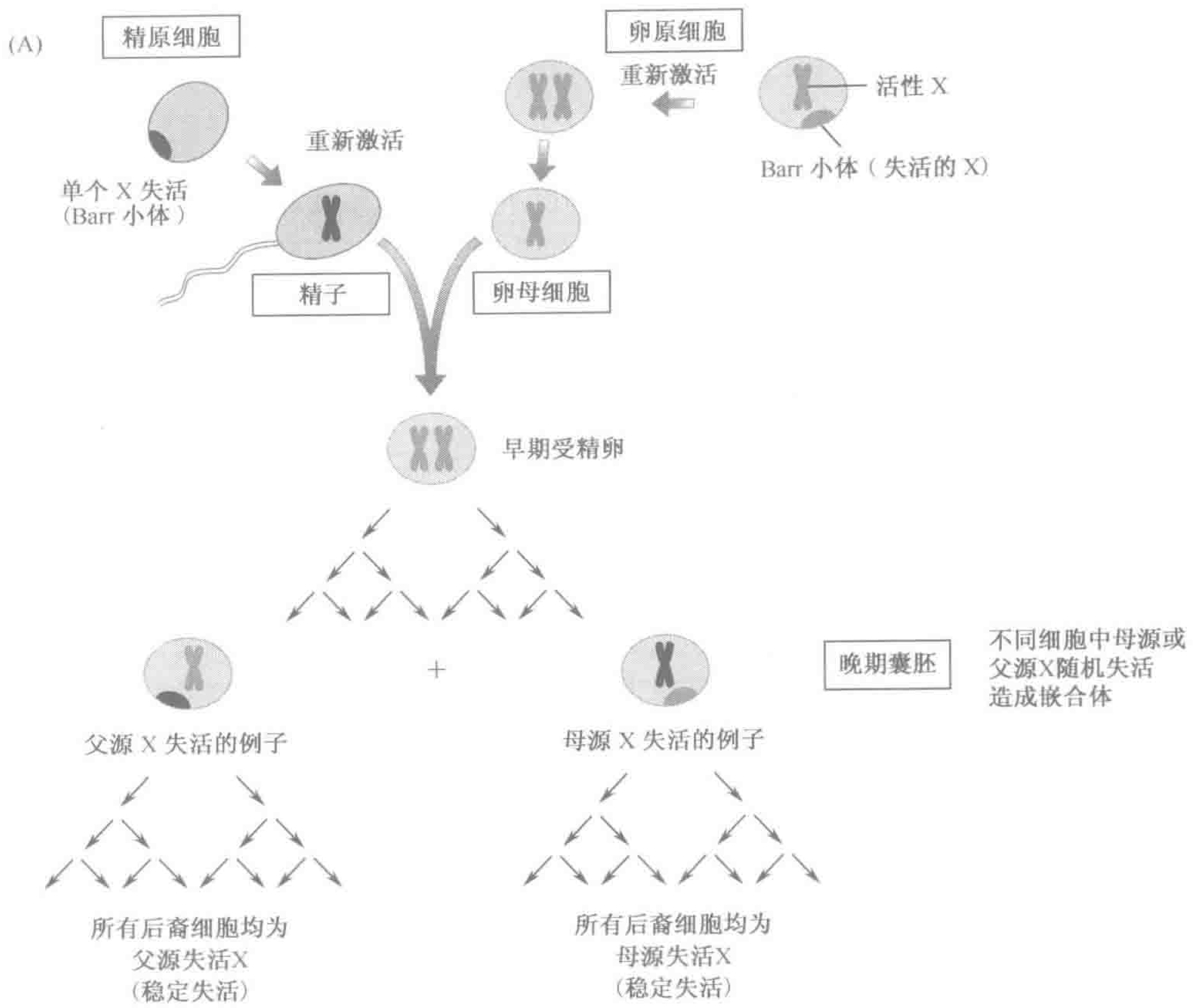
X 染色体失活的过程很复杂，而截然不同的分子机制将参与失活的起始以及失活的维持（Avner and Heard，2001；Brockdorff，2001）。在遗传学上已经明确两个重要的顺式作用元件：

- X **失活中心**（X-inactivation center，Xic），控制 X 失活的起始和扩延。*Xic* 的存在对于 X 失活至关重要：携带 Xic 的 X 染色体可以遭受失活，而缺乏它的则不能。*Xic* 对于染色体‘计数’也很重要。在少数具有异常数目的 X 染色体（45，X；47，XXX；47，XXY 等）的个体中，无论存在多少条 X 染色体，只有一条将保持活性。相比之下，在三倍体个体中，一条或两条 X 染色体将保持活性，而在四倍体个体中，两条 X 染色体将保持活性。某种计数机制将确保每两套常染色体会有一条 X 染色体保持活性。*Xic* 的功能依赖于 *XIST*，一个编码一种功能性非编码 RNA 的基因（见下文），尽管 *XIST*（*xist*）对于启始 X 染色体失活至关重要，但它在维持失活状态中却并非必需；
- X **调控元件**（X-controlling element，*Xce*），它将影响对于哪条 X 染色体会保持活性，哪条会失活的选择。X 失活的倾斜（由失活与活性 X 在正常情况下的 50：50）可发生于 *Xce* 等位基因为杂合性的女性中（含有强 *Xce* 等位基因的 X 染色体比含有弱 *Xce* 等位基因者可能更具有活性）。*Xce* 与 *Xic* 和 *Xist* 不同，并定位于它们的端粒侧，但其分子基础仍不清楚。

X 失活中心定位于人类的 Xq13 之后，对该区域的分析揭示了一个特别的基因：*XIST*（在啮齿类中称作 *Xist*）。*XIST* 呈单等位性表达，并独特地表达自失活的 X 染色体。它的原始转录物将经历剪接和多聚腺苷酸化从而产生一个 17 kb 的成熟非编码 RNA。它似乎是沿合成它的 X 染色体传播这种转录失活状态的主要信号，而不知何故，这种顺式有限的 RNA 产物的扩延发挥作用以至于将跨越很长的距离失活的 X 染色体包裹起来。而后这种 RNA 被认为可募集一些蛋白质因子将染色质组织成为一种封闭的转录失活的构象（Brockdorff，2002）。

另一个特别的基因 *TSIX*（在啮齿类中称作 *TsiX*）具有一个与整个 *XIST* 基因相重叠的转录单位，但却位于反义链上。这个伴侣基因表达于未分化的胚胎干细胞和早期胚胎中，并推测在 X 失活的起始中顺式调控 *XIST* 基因的表达（Avner and Heard，2001；Brockdorff，2002）。

(A)



(B)

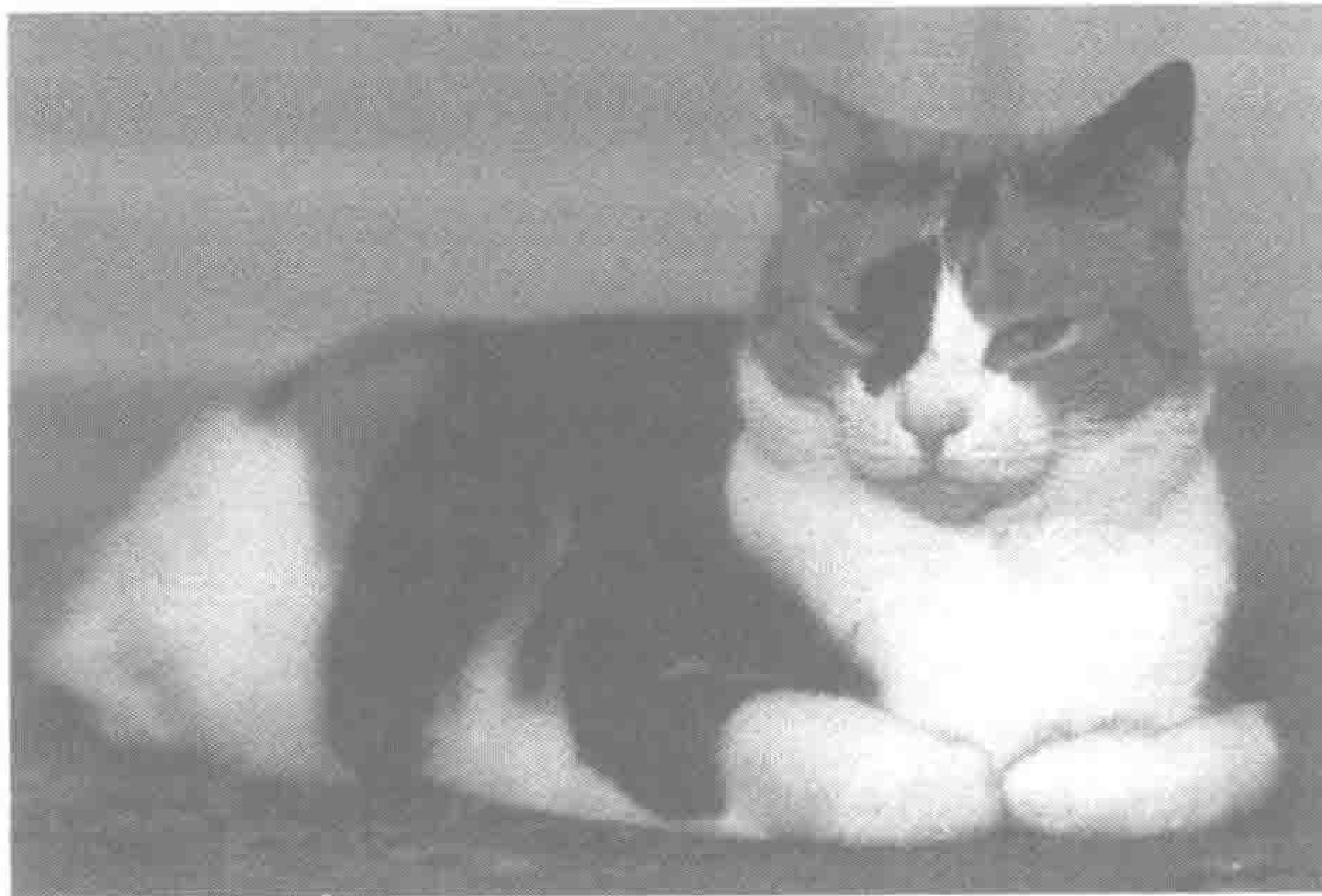


图 10.26　哺乳动物中 X 染色体的失活

(A) **X 失活的过程**。在早期的 XX 女性受精卵中，两条 X 染色体均有活性，但大约在晚囊胚期阶段在每个细胞中将随机选择失活一条父源或母源 X，一个细胞的选择将在其所有的子细胞中被保存。成年 XX 女性拥有父源或母源 X 失活的细胞的克隆群。失活的 X 将在减数分裂之前的某个时间在卵母细胞中重新激活。在精子发生过程中，X 和 Y 染色体均短暂的失活。经 Elsevier Trends 杂志允许，改编自 Migeon(1994). Trends Genet. 10，230～235。(B) **三色（龟甲纹及白色）猫**。这种猫（总是雌性）具有一条含有编码黑色皮毛基因的 X 染色体和另一条含有编码橙色皮毛基因的 X 染色体。产生不同颜色的皮毛斑块是由于克隆化的 X 失活所致。白色皮毛斑块是由另一个基因决定。

10.6 Ig 和 TCR 基因的特殊结构与表达

如图 9.10 所示，免疫球蛋白（Ig）和 T 细胞受体（TCR）为异二聚体。在人类细胞中有三种 Ig 基因：一种 *IGH* 编码重链，*IGK* 和 *IGC* 两种基因则分别编码可相互替换的 κ 和 λ 轻链。四种人类 TCR 基因分别编码四种不同的 TCR 链：形成常见的 αβ 异二聚体的 α 和 β，加上形成较为罕见的 γδ 异二聚体的 γ 和 δ。

这七种基因的结构和表达在许多方面与其他基因大不相同。之所以如此，是因为每个人都需要产生巨大数量（达到大约 2.5×10^7 不同种类的数量）分别由 B 和 T 淋巴细胞所产生的 Ig 和 TCR。这些细胞是**适应性免疫系统**（adaptive immune system）的主要因子，而它们需要识别无数的外来抗原，从而能够装备对于大量不同病原体的抵抗。然而，单个的 B 或 T 细胞却是**单一特异性**（monospecific）的：它将产生单一类型的 Ig 或 TCR 异二聚体，具有一个独特的抗原结合位点，因而对特定的抗原具有特异性。在任何一个个体中，是各种各样 B 或 T 细胞的总体使合成如此多不同类型的这些分子成为可能。通过提供一个巨大的 Ig 和 TCR 种类储备，大大提高了识别并结合很多不同类型的外来抗原的可能性。

Ig 和 TCR 基因在大多数细胞中是无活性的，但这些基因的非凡混编将分别发生于 B 和 T 细胞中以激活它们。当这种情况发生时，将产生新的编码组合，使一个个体能够产生种类惊人的 Ig 和 TCR。四种主要的 DNA 混编机制据知作用于脊椎动物中：

- ▶ **VDJ 或 VJ 重组，一种广泛的机制**。Ig 和 TCR 的可变区均由两种或三种基因片段编码，始终是一个 V（可变区）和一个 J（连接区）基因片段，而在某些情况下还有一个 D（差异区）基因片段。在种系 DNA 中，每个这种基因片段均存在多个拷贝（表 10.8 和节 10.6.1）。然而，成熟的 B 细胞和 T 细胞的 DNA 将经历细胞特异性 DNA 重组，使得 V、D 和 J 片段，或一个 V 和 J 片段集合形成一个特定组合，分别产生一个 VDJ 外显子或一个 VJ 外显子（节 10.6.1）；

表 10.8　人类的 Ig 和 TCR 基因

基因	位置	V,D,J 基因片段的数目			C 转录单位的数目
		V	D	J	
IGH	14q32.3	123-129[a]	27	9	11
IGA	2p12	76	0	5	1
IGL	22q11	70－71[a]	0	7－11[a]	7－11[a]
TRA	14q11.2	49(＋5)[b]	0	61	1
TRB	7q34	64－67[a]	2	14	2
TRG	7p15-p14	12－15[a]	0	5	2
TRD	14q11.2	1(＋5)[b]	3	4	1

注：数目包括非功能性（假基因）序列——一个来自 *IGH* 基因的例子见图 10.27。

a 不同单体型上的数目不同。b 五个 V 基因片段为相邻的 *TRA* 和 *TRD* 基因所共有。图示见不同数据库如 Atlas of Genetics and Cytogenetics in Oncology and Haematology，网址 www.infobiogen.fr/services/chromcancer/Genes/Geneliste.html 中单个基因名字下面的内容。

- ▶ **体细胞高突变，人和小鼠中的一种主要机制**。当点突变因易于出错的 DNA 修复被

引入 VJ 及 VDJ 外显子的 V 区时，将产生额外的 Ig 多样性。

- **基因转变，家兔和鸡中的一种主要机制**。当成段的核苷酸序列从上游的 *V* 假基因（*ψV*）节段拷贝至 VJ 和 VDJ 外显子中的 V 基因节段时，亦将产生 Ig 多样性。
- **类型切换重组，一种广泛的机制**。Ig 的恒定区由包含几个外显子的转录单位编码。对于 Ig 重链而言，几种不同的转录单位将编码不同种类的重链，例如 Cμ(IgM)，Cδ(IgD)，Cγ(IgG) 等。在 B 细胞的成熟过程中，重链基因内部的染色单体内重组将使特定种类的恒定区转录单位到达 VDJ 外显子的附近（节 10.6.2）。

最近，已证实体细胞高突变，基因转位和类型切换重组活动（它们可全体发生于一个物种中，尽管一些机制在特定的物种中尤为常见）全体由一个基因调控，后者将编码**激活诱导的脱氧胞嘧啶脱氨基酶**（activation-induced deoxycytidine deaminase，AID；Petersen-Mahrt *et al*.,2002）。

10.6.1　B 和 T 细胞中的 DNA 重排将产生编码 Ig 和 TCR 可变区的细胞特异性外显子

编码三种 Ig 链（一种重链和两种轻链）以及四种 TCR 链（α、β、γ 和 δ）的基因位于不同的染色体，并显示一种特别的结构。在每种情况下，可变区由两到三个不同的基因片段编码，后者则呈现为连续重复于种系 DNA 中众多不同的拷贝（表 10.8 和图 10.27）。这些基因片段包括：

- V（**可变区**）**基因片段** [V(variable region) gene segment]。V 基因片段编码大多数可变区；
- J（**连接区**）**基因片段** [J(joining region) gene segment]。J 基因片段编码连接区、可变区 C 末端的一小部分；
- D（**差异区**）**基因片段** [D(diversity region) gene segment]。D 基因片段编码 Ig 重链、TCRβ 链以及 TCRδ 链的可变区 C 端附近的一个小的**差异区**（diversity region）。

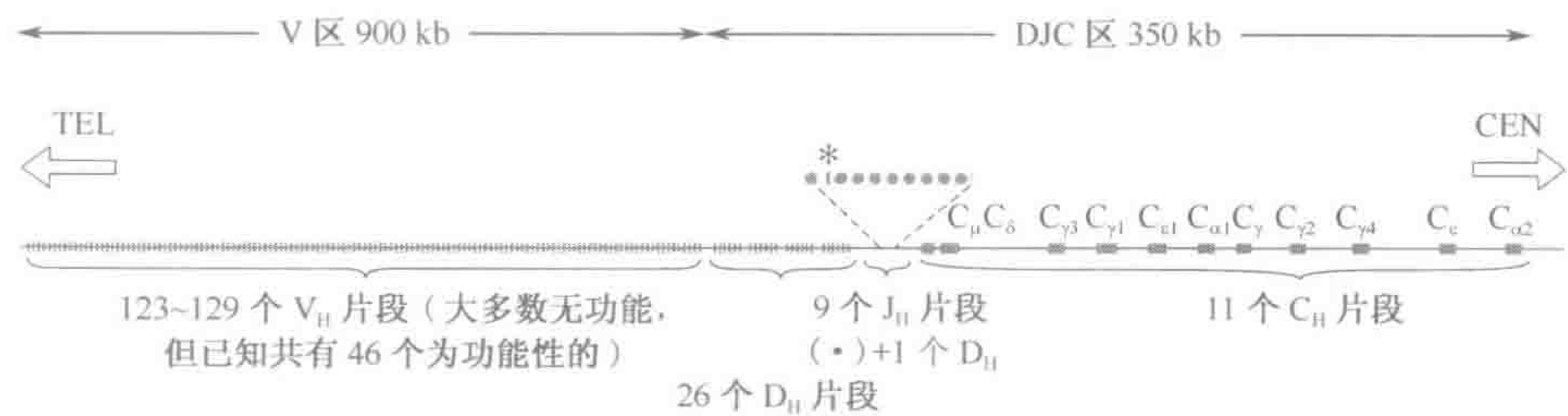


图 10.27　位于 14q32 的 *IGH* Ig 重链基因的多个基因片段

该基因跨越 14q32.3 从端粒末端（左）到着丝粒末端（右）1250 kb 的区域。有 123～129 个 V_H 基因片段（取决于单体型），其中大多数据知为无功能的假基因片段，27 个 D_H 基因片段、9 个 J_H 基因片段和 11 个 C 转录单位（各包含若干外显子）。更多的细节见 http://www.infobiogen.fr/services/chromcancer/Genes/IgHID40.html。

Ig 和 TCR 基因簇内基因片段的独特排列反映了 B 和 T 淋巴细胞经过体细胞重组来激活原来无功能的 Ig 和 TCR 基因并随后自它们表达功能性产物的很不寻常的方式。它们通过聚集（V+D+J）基因片段（对于编码 Ig 重链、TCRβ 和 TCRδ 基因而言）或 V+J 基因片段（对于其他四种基因而言）的特殊组合来实现这一过程。一经组合，V+J

或 V+D+J 单位将作为功能性外显子而运转（见下文）。对于这七种基因的每一种而言，对于将许多 V，D 或 J 基因片段中的哪些凑到一起的选择将因淋巴细胞而异，因此新产生的 VJ 外显子和 VDJ 外显子将为细胞特异性（图 10.28）。因此，个体 B 和 T 淋巴细胞将产生不同的 Ig 和 TCR。因此，在某种意义上，就 B 和 T 淋巴细胞中的 Ig 和 TCR 基因的结构而言，每个人都是一个嵌合体，并且甚至同卵孪生子也将在遗传学上歧化。

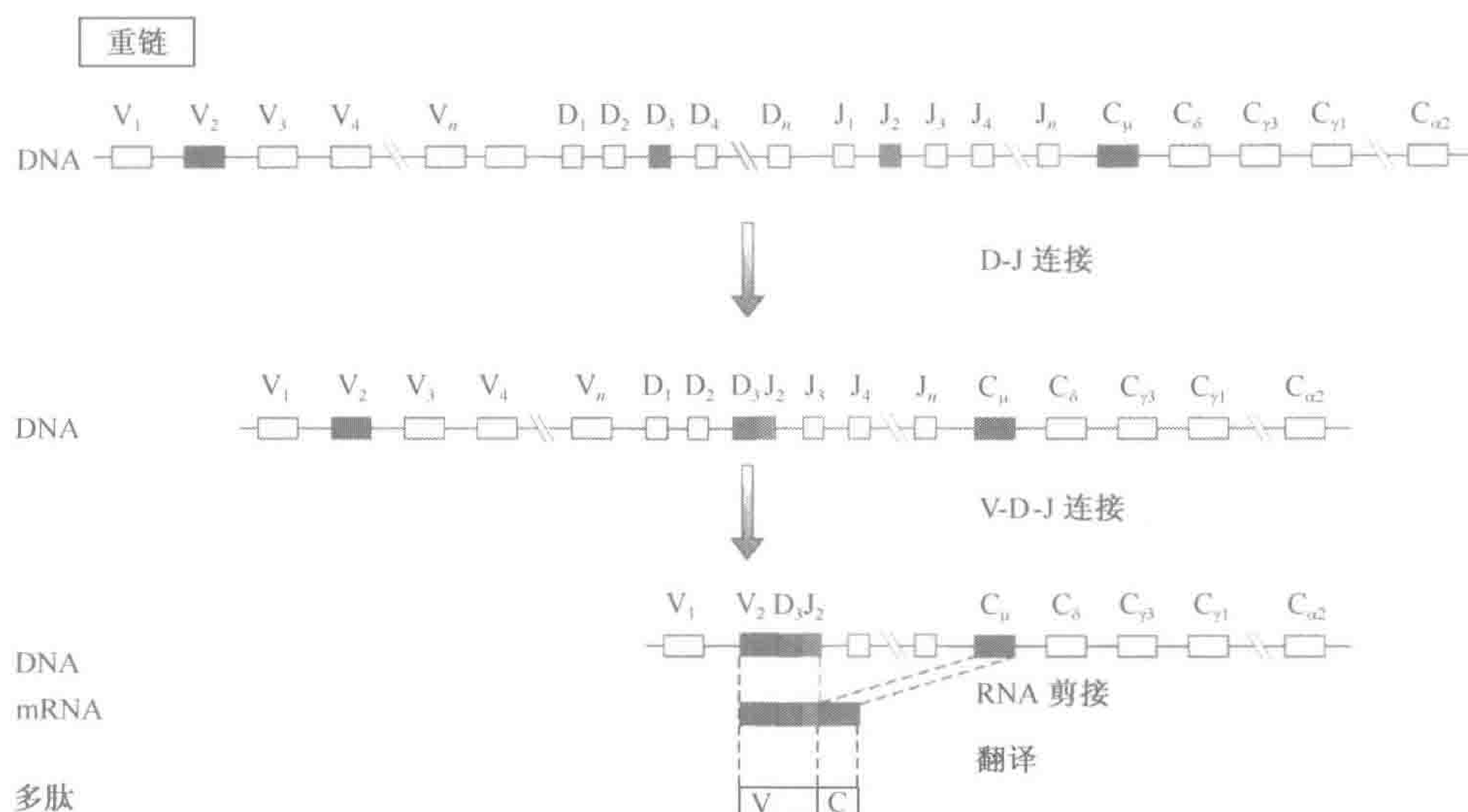


图 10.28　细胞特异性的 VDJ 重组作为产生 Ig 重链的开头

两次连续的体细胞重组将首先产生第一个 D-J 连接，之后是一个成熟的 VDJ 外显子。在这个特别的例子中，129 个不同的 V 片段中的第二个（V_2）将融合到第三个 D 区域片段（D_3）以及第二个 J 区域片段（J_2）以产生一个功能性 $V_2D_3J_2$ 外显子，但该选择为细胞特异性，因此相邻的 B 淋巴细胞可能有一个譬如 $V_{129}D_{17}J_1$ 的功能性外显子。一旦 VDJ 外显子已组合完成，该基因即可转录，以这个 VDJ 外显子作为第一外显子，而随后的外显子将由最近的 C 转录单位提供。首先，Cμ 和 Cδ 转录单位是最近的，而最先产生的 Ig 重链类型将为 μ 链，然后是 μ 加上 δ，分别为具有 IgM 和 IgD 特征的重链。然而，随着 B 细胞的成熟，随后的体细胞重组将导致原先组合的 VDJ 外显子连接到不同的 C 转录单位上（重链类型切换，见正文及图 10.29）。

一旦某个 VDJ 或 VJ 外显子已组合完毕，一个功能性的 Ig 或 TCR 基因即已形成。此时新的 VDJ 或 VJ 外显子将为这个基因提供第一个外显子，而下游的外显子则将由邻近的 C 转录单位中的外显子提供。对于 Ig 重链基因而言，存在若干具有不同生物学特性的各不相同的功能性 C 转录单位。然而，最初剪接将涉及 VDJ 外显子以及邻近的 Cμ 和 Cδ 转录单位的外显子，而选择性剪接则可导致 μ 和 δ 链的合成，后者将被并入 IgM 和 IgD 免疫球蛋白中去（图 10.28）。然而，随着 B 细胞的成熟，随后的体细胞重组将使原先组合的 VDJ 外显子连接到不同的 C 转录单位上以产生将被并入 IgG、IgA 或 IgE 中去的 γ、α 或 ε 重链［重链**类型切换**（class switch），见正文及图 10.29］。

导致产生功能性 VJ 和 VDJ 外显子的遗传机制常常涉及分隔所选基因片段的序列的大范围缺失（最有可能通过染色单体内重组，与图 10.29B 中的方式极为相似）以及在某些情况下居间序列的倒位。保守的**重组信号序列**（recombination signal se-

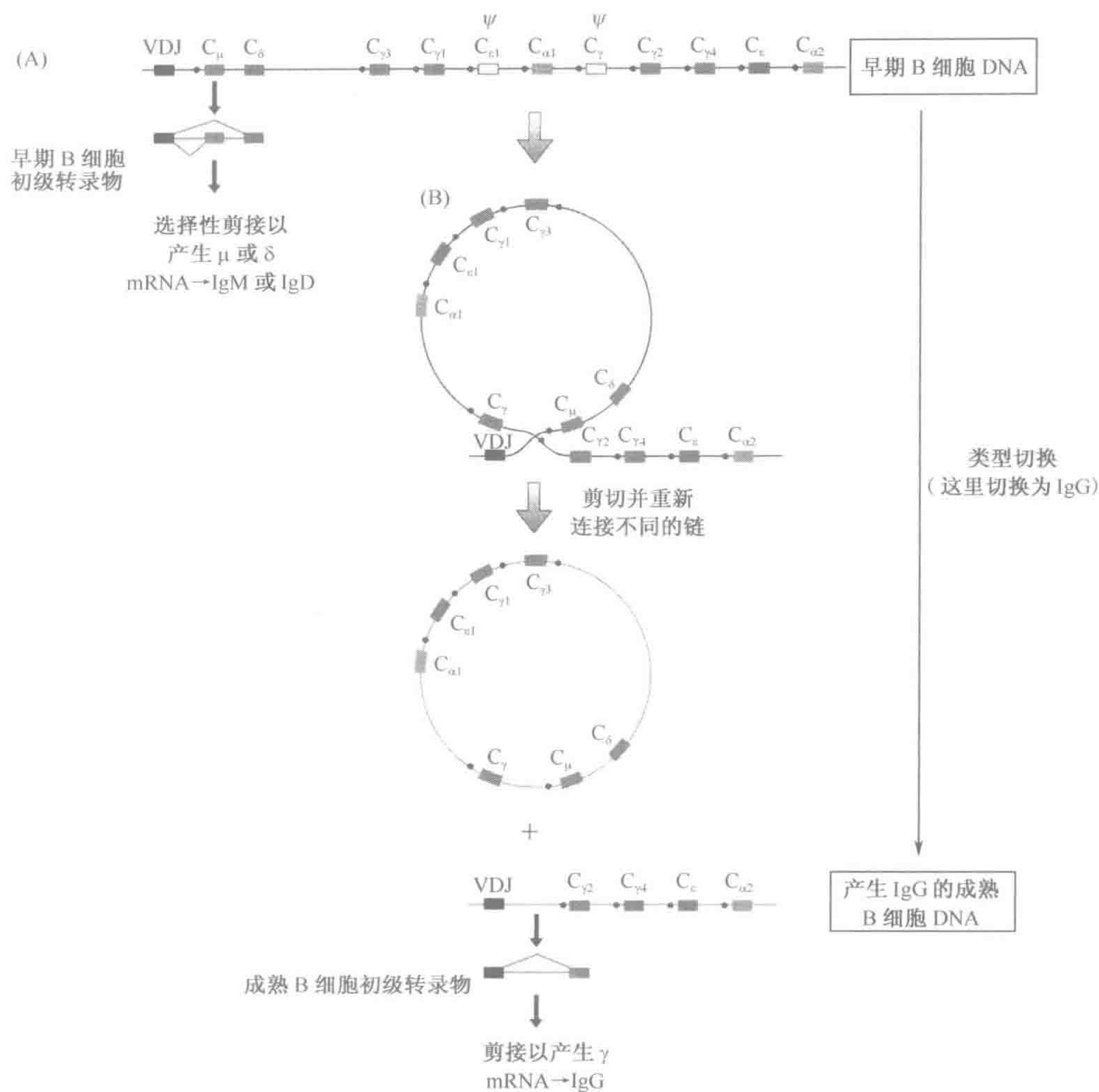


图 10.29　Ig 重链类型切换由染色单体内重组介导

(A) 向 IgD **的早期（部分）切换**。最初的重链类型为 IgM，因为 RNA 剪接将转录自 VDJ 外显子和邻近的 Cμ 转录单位的外显子的序列结合到一起。然而，随着 B 细胞的成熟，选择性 RNA 剪接将转录自 VDJ 外显子和 Cδ 转录单位的外显子的序列结合到一起，导致产生额外的 IgD。(B) **向** IgA、IgG **和** IgE **的后期切换**。在这里类型切换将通过染色单体内重组而发生，其中相同的 VDJ 外显子被拉近到最初更远侧的恒定区转录单位：一个 Cα、Cγ（如这里所示）或 Cε 转录单位。新的 VDJ-C 组合将表达产生 IgA、IgG 或 IgE。

quence）位于每个 V 和 J 片段的 3′端以及每个 D 基因片段 5′和 3′端的旁侧。它们可使 V 链连接到 J，或者 D 连接到 J，随后是 V 连接到 DJ，但绝不会使 V 连接到 V 或者 D 连接到 D 等。

这种重组活动将由 V（D）J **重组酶**（recombinase），一种含有两种淋巴细胞特异性蛋白质 RAGI 和 RAG2 的复合体以及在我们所有的细胞中帮助修复受损 DNA 的酶来指引。RAGI 和 RAG2 在重组信号序列处造成双链断裂，而这种断裂将通过缝缀恰当的 V、D 以及 J 基因片段，但排除间隔序列而修复。尽管位点特异性重组通常是精确的，

但分别发生于 B 和 T 细胞中 Ig 和 TCR 基因处的重组却故意地不精确。相反，不同数目的核苷酸常常自正在重组的基因片段的末端丢失，并且一个或更多随机选择的核苷酸也可能被插入，形成**接合多样性**（junctional diversification）。这就极大地增加了可变区编码序列的多样性。

10.6.2　重链类型切换涉及将单个 VDJ 外显子与选择性恒定区转录单位相连接

尽管一个 B 细胞只能产生一种类型的 Ig 分子，重链的类型（或同型）在发育过程中可发生变化：在由一个原始细胞所产生的单一谱系后裔细胞中，可产生一种具有与原先一样的抗原结合位点的 Ig，但却使用不同类型的重链［类型切换（class switching）或**同型切换**（isotype switching）］。这种切换涉及将两次连续的体细胞重组（图 10.28）结合到一起的相同 VDJ 外显子有区别地连接到选择性恒定区转录单位上。

类型切换涉及染色单体内重组，后者将导致将一个 VDJ 外显子连接到一个更远侧的恒定区转录单位上（VDJ-C **连接**，VDJ-C joining）。它将涉及以下的进展：

▶（Ⅰ）**由未成熟的 B 稚细胞完成的仅有 IgM 的初步合成**。这发生在 B 细胞发育的早期，因为 RNA 剪接将使转录自 VDJ 外显子和邻近的 $C\mu$ 转录单位的外显子的序列凑到一起（图 10.29A）；

▶（Ⅱ）**由成熟的 B 稚细胞完成的 IgM 和 IgD 的后期合成**。在 B 细胞发育稍晚，但仍处于**免疫幼稚**（immunologically naive）阶段（即它们还未被暴露于外源性抗原）时，将发生一次部分类型切换，而 B 细胞这时将产生 IgD 以及 IgM。这是由于此时额外的可变 RNA 剪接也可使转录自 VDJ 外显子和相邻的 $C\delta$ 转录单位的外显子的序列结合到一起（图 10.29A）。大多数的 IgM 和 IgD 产物被专门用于制造膜结合受体，但暴露于外源性抗原将会触发在一次微弱的**初次免疫应答**（primary immune response）中分泌可溶性的 IgM 抗体；

▶（Ⅲ）**由成熟 B 细胞完成的 IgM、IgE 或 IgA 的合成**。暴露于外源性抗原之后，B 细胞将出现改善 Ig 结合的亲和力的作用（**亲和力成熟**，affinity maturation），因此它们能够在未来的场合对外源性抗原作出更有效的反应［那时它们将会在强大的**二次免疫应答**（secondary immune response）中针对外源性抗原分泌大量具有极高亲和力的可溶性抗体］。亲和力成熟将受助于体细胞超突变/基因转变事件等。在成熟的 B 细胞中，此时将通过一种不同的机制发生类型切换并涉及一种重组活动，在 DNA 水平使同样的 VDJ 外显子连接到 $C\gamma$、$C\varepsilon$ 或 $C\alpha$ 转录单位。这种机制涉及通过染色单体内重组（图 10.29B）造成间隔序列的缺失。成熟 B 细胞分泌的 Ig 分子能够结合至不同的细胞类型，后者具有这些可溶性抗体尾部（Fc）的特化受体。这种 Fc 受体能够选择性地结合不同种类的 Ig：

- IgG——除激活**补体系统**（complement system）外，IgG 能够被**巨噬细胞**（macrophage）和**中性粒细胞**（neutrophil）等强大的吞噬细胞上的 Fc 受体特异性结合；
- IgA——一种仅分泌性上皮才有的 Fc 受体，能转运 IgA 抗体，因此它们将构成分泌物，包括唾液、泪液、乳汁以及呼吸系统及肠道的分泌物中主要的抗体类型；
- IgE——**肥大细胞**（mast cell）和**嗜碱性粒细胞**（basophil）上的 Fc 受体与 IgE 以

极高的亲和力结合。结合的 IgE 分子随后将作为被动获得的抗原受体而发挥作用。随后抗原的结合将触发肥大细胞或嗜碱性粒细胞分泌多种细胞因子及组胺。另外，肥大细胞还将分泌一些能吸引并激活**嗜酸性粒细胞**（eosinophil）的因子，后者亦具有能结合 IgE 分子的 Fc 受体并能够杀死各种类型的寄生虫。

10.6.3　Ig 和 TCR 的单一特异性是由于等位基因及轻链排斥所致的

在人类细胞中有三种功能性 Ig 基因（编码重链的 *IGH*，以及两种编码在功能上可互换的 κ 及 λ 轻链的 *IGK* 和 *IGL* 基因），并且由于这些基因同时存在于母源及父源同源体上，因此共有六个基因可能被用于产生 Ig 链。然而，单个 B 细胞是**单一特异性**（monospecific）的：它只能产生单一类型的 Ig 分子，具有单一类型的重链以及单一类型的轻链。这是出于两种原因：

- **等位基因排斥**（allelic exclusion）。在任何一个 B 细胞中一条轻链或一条重链可从一条母源染色体或一条父源染色体合成，但不能从两条亲代同源体合成。因此，在 B 细胞中的重链基因座上存在单等位性表达。这种现象亦牵涉到 TCR 基因簇；
- **轻链排斥**（light chain exclusion）。在单个 B 细胞中合成的轻链可以是 κ 链或 λ 链，但绝不会两者兼有。作为这一必备条件加上等位基因排斥的结果，在两个功能性轻链基因簇之一上存在单等位性表达，而另一个则无表达。

对于两个重链等位基因中的哪一个来合成重链以及四个潜在的轻链等位基因中的哪一个来合成轻链的选择似乎是随机的。最可能的情况是，在每个 B 细胞前体中，生产性的 DNA 重排曾尝试于全部六个 Ig 基因上，然而在多于一个轻链基因或多于一个重链基因中产生生产性重排的机会则不大。然而，另外似乎也存在某种**负反馈调节**（negative feedback regulation）：一个重链等位基因上的功能性重排将抑制发生在另一个等位基因上的重排，而能够编码轻链的四个基因中的任何一个的功能性重排将抑制发生在其他三个基因上的重排。

（李英慧　译）

进一步阅读

Brivanou AH, Darnewell Jr. JE (2002) Signal transduction and the control of gene expression. *Science* **295**, 813–818.

Gellert M (2002) V(D)J recombination: RAG proteins, repair factors and regulation. *Annu. Rev. Biochem.* **71**, 101–132.

Imprinted Gene Catalogue at http://cancer.otago.ac.nz/IGC/Web/home.html

Janeway CA, Travers P, Walport M, Capra JD (2001) *Immunobiology. The Immune System in Health and Disease.* 5th Edn. Garland Publishing. Available with animations at http://www.blink.uk.com/immunoanimations/

Latchman D (1998) *Gene regulation. A Eukaryotic Perspective.* Stanley Thornes (Publishers) Ltd., Cheltenham.

Orphanides G, Reinberg D (2002) A unified theory of gene expression. *Cell* **108**, 439–451.

Palacois IM, St. Johnson D (2001) Getting the message across: the intracellular localization of mRNAs in higher eukaryotes. *Annu. Rev. Cell Dev. Biol.* **17**, 569–614.

Plath K, Mylnarczyk-Evans S, Nusinow DA, Panning B (2002) *XIST* RNA and the mechanism of X chromosome inactivation. *Annu. Rev. Genet.* **36**, 233–278.

Sandberg K, Mulroney SE (eds) (2002) *RNA Binding Proteins: New Concepts in Gene Regulation.* Kluwer Academic Publishers, Boston.

Turner BM, Turner BS (2002) *Chromatin and Gene Regulation: Mechanisms in Epigenetics.* Blackwell Science, Oxford.

Travers A (1993) *DNA–Protein Interactions.* Chapman & Hall, London.

van Driel R, Otte AP (1997) *Nuclear Organization, Chromatin Structure and Gene Expression.* Oxford University Press, Oxford.

参考文献

Avner P, Heard E (2001) X-chromosome inactivation: counting, choice and initiation. *Nature Rev. Genet.* **2**, 59–67.

Ayoubi TA, Van De Ven WJ (1996) Regulation of gene expression by alternative promoters. *FASEB J* **10**, 453–460.

Bell AC, West AG, Felsenfeld G (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science* **291**, 447–450.

Bannerjee D, Slack F (2002) Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *BioEssays* **24**, 119–128.

Berger SL (2002) Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.* **12**, 142–148.

Bestor TH (1998) Cytosine methylation and the unequal developmental potentials of the oocyte and sperm genomes. *Am. J. Hum. Genet.* **62**, 1269–1273.

Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21.

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110.

Brockdorff N (2002) X chromosome inactivation: closing in on proteins that bind *Xist* RNA. *Trends Genet.* **18**, 352–358.

Bulger M, Groudine M (1999) Looping versus linking: toward a model for long-distance gene activation. *Genes Dev.* **13**, 2465–2477.

Bulger M, Sawado T, Schubeler D, Groudine M (2002) ChIPs of the β-globin locus: unravelling gene regulation within an active domain. *Curr. Opin. Genet. Dev.* **12**, 170–177.

Butler JE, Kadonaga JT (2002) The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* **16**, 2583–2592.

Caceres JF, Kornblihtt AR (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends Genet.* **18**, 186–193.

Chess A (1998) Expansion of the allelic exclusion principle. *Science* **279**, 2067–2068.

De Moor CH, Richter JD (2001) Translational control in vertebrate development. *Int. Rev. Cytol.* **203**, 567–608.

Dever TE (2002) Gene-specific regulation by general translation factors. *Cell* **108**, 545–556.

Edwards-Gilbert G, Veraldi KL, Milcarek C (1997) Alternative poly (A) site selection in complex transcription units: means to an end? *Nucl. Acid Res.* **13**, 2547–2561.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013.

Finta C, Zaphiropoulos PG (2002) Intergenic mRNA molecules resulting from trans-splicing. *J. Biol. Chem.* **277**, 5882–5890.

Fuks F, Hurd PJ, Wolf D, Nan X, Bird AP, Kouzarides T (2002) The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation. *J. Biol. Chem* **278**, 4035–4040.

Gallagher RC, Pils B, Albalwi M, Francke U (2002) Evidence for the role of PWCR1/HBII-85 C/D box small nucleolar RNAs in Prader-Willi syndrome. *Am. J. Hum. Genet.* **71**, 669–678.

Gerber AP, Keller W (2001) RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends Biochem. Sci.* **26**, 376–384.

Goll MG, Bestor TH (2002) Histone modification and replacement in chromatin activation. *Genes Dev.* **16**, 1739–1742.

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107.

Gray NK, Wickens M (1998) Control of translation initiation in animals. *Annu. Rev. Cell Dev. Biol.* **14**, 399–458.

Hazelrigg T (1998) The destinies and destinations of RNAs. *Cell* **95**, 451–460.

Jones PA (1999) The DNA methylation paradox. *Trends Genet.* **15**, 34–37.

Karin M, Hunter T (1995) Transcriptional control by protein phosphorylation: signal transmission from the cell surface to the nucleus. *Curr. Biol.* **5**, 747–757.

Klausner RD, Rouault TA, Harford JB (1993) Regulating the fate of mRNA: the control of cellular iron metabolism. *Cell* **72**, 19–28.

Kleinjan D-J, van Heyningen V (1998) Position effects in human genetic disease. *Hum. Molec. Genet.* **7**, 1611–1618.

Larsson SH, Charlieu JP, Miyagawa K *et al.* (1995) Subnuclear localization of WT1 in splicing or transcription factor domains is regulated by alternative splicing. *Cell* **81**, 391–401.

Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.* **14**, 2551–2569.

Li E (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Rev. Genet.* **3**, 662–673.

Lopez AJ (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. *Annu. Rev. Genet.* **32**, 279–305.

Lunyak VV, Burgess R, Prefontaine GG *et al.* (2002) Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science* **298**, 1747–1752.

Lyon MF (1999) X-chromosome inactivation. *Curr. Biol.* **9**, R235–R237.

Maher E, Reik W (2000) Beckwith-Wiedemann syndrome: imprinting in clusters revisited. *J. Clin. Invest.* **105**, 247–252.

Mahmoudi T, Verrijzer CP (2001) Chromatin silencing and activation by Polycomb and trithorax group proteins. *Oncogene* **20**, 3055–3066.

Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418**, 236–243.

Martin DIK (2001) Transcriptional enhancers – on/off gene regulation as an adaptation to silencing in higher eukaryotic nuclei. *Trends Genet.* **17**, 444–448.

Meguro M, Kashiwagi A, Mitsuya K, Nakao M, Kondo I, Saitoh S, Oshimura M (2001) A novel maternally expressed gene, ATP10C, encodes a putative aminophospholipid translocase associated with Angelman syndrome. *Nature Genet.* **28**, 19–20.

Merteneit C, Yoder JA, Taketo T, Laird DW, Trasier JM, Bestor TH (1998) Sex-specific exons control DNA methyltransferase in mammalian germ cells. *Development* **125**, 889–897.

Migeon BR (1994) X-chromosome inactivation: molecular mechanisms and genetic consequences. *Trends Genet.* **10**, 230–235.

Narlikar GJ, Fan H-Y, Kingston RE (2002) Co-operation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475–487.

Ogbourne S, Antalis TM (1998) Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem. J.* **331**, 1–14.

Ohlsson R, Tycko B, Sapienza C (1998) Monoallelic expression: 'there can only be one'. *Trends Genet.* **14**, 435–438.

Pasquinelli AE, Ruvkun G (2002) Control of developmental timing by microRNAs and their targets. *Annu. Rev. Cell. Dev. Biol.* **18**, 495–513.

Pekhletsky RI, Chernov BK, Rubtsov PM (1992) Variants of the 5′-untranslated sequence of human growth hormone receptor mRNA. *Mol. Cell. Endocrinol.* **90**, 103–109.

Perk J, Makedonski K, Lande L, Cedar H, Razin A, Shemer R (2002) The imprinting mechanism of the Prader-Willi/Angelman regional control center. *EMBO J.* **21**, 5807–5814.

Petersen-Mahrt SK, Harris RS, Neuberger MS (2002) AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–10.

Razin A, Kafri T (1994) DNA methylation from embryo to adult.

Prog. Nucl. Acid Res. Mol. Biol. **48**, 53–81.

Reik W, Dean W, Walter J (2001) Epigenetic reprogramming in mammalian development. *Science* **293**, 1089–1093.

Roberts GC, Smith CWJ (2002) Alternative splicing: combinatorial output from the genome. *Curr. Opin. Chem. Biol.* **6**, 375–383.

Rougeulle C, Heard E (2002) Antisense RNA in imprinting: spreading silence through *Air*. *Trends Genet.* **18**, 434–437.

Runte M, Huttenhofer A, Gross S, Keifmann M, Horsthemke B, Buiting K (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genet.* **10**, 2687–2700.

Schmucker D, Clemens JC, Shu H *et al.* (2000) *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**, 671–684.

Schoenherr CJ, Paquette AJ, Anderson DJ (1996) Identification of potential target genes for the neuron-restrictive silencer factor. *Proc. Natl Acad. Sci. USA* **93**, 9881–9886.

Simmen MW, Leitgeb S, Charlton J, Jones SJM, Harris B, Clark VH, Bird AP (1999) Nonmethylated transposable elements and methylated genes in a chordate genome. *Science* **283**, 1164–1167.

Siomi H, Dreyfuss G (1997) RNA-binding proteins as regulators of gene expression. *Curr. Opin. Genet. Dev.* **7**, 345–353.

Skuse DH, James RS, Bishop DV *et al.* (1997) Evidence from Turner's syndrome of an imprinted X-linked locus affecting cognitive function. *Nature* **387**, 705–708.

Sleutels F, Barlow DP (2002) The origins of genomic imprinting in mammals. *Adv. Genet.* **46**, 119–163.

Sleutels F, Zwart R, Barlow DP (2002) The non-coding *Air* RNA is required for silencing autosomal imprinted genes. *Nature* **415**, 810–813.

Turner BM (2002) Cellular memory and the histone code. *Cell* **111**, 285–291.

Tycko B, Morison IM (2002) Physiological functions of imprinted genes. *J. Cell Physiol.* **192**, 245–258.

Wahle E, Kuhn (1997) The mechanism of 3′ cleavage and polyadenylation of 3′ eukaryotic pre-mRNA. *Prog. Nucl. Acid Res. Mol. Biol.* **57**, 41–71.

Walsh CP, Bestor TH (1999) Cytosine methylation and mammalian development. *Genes Dev.* **13**, 26–34.

Wickens M, Anderson P, Jackson RJ (1997) Life and death in the cytoplasm: messages from the 39′ end. *Curr. Opin. Genet. Dev.* **7**, 233–241.

Wickens M, Bernstein DS, Kimble J, Parker RA (2002) PUF family portrait: 3′ UTR regulation as a way of life. *Trends Genet.* **18**, 150–157.

Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340.

第 11 章　人类基因组的不稳定性：突变与 DNA 复制

本章内容

11.1　突变、多态性与 DNA 修复概述

同其他基因组一样，人类基因组的 DNA 并非是一个静态实体。相反，它将遭受各种不同类型的可遗传改变（突变）。大范围的染色体畸变涉及染色体的丢失或获得，或者染色质的断裂与重接（节 2.5）。根据对 DNA 序列的影响，小范围的突变可分为下列不同的突变类型：

- **碱基置换**（base substitution）——涉及通常为单个碱基的替换；少数情况下几个成簇的碱基也可能通过一种形式的基因转变同时被替换（节 11.3.3）。
- **缺失**（deletion）——一个或多个核苷酸从序列中删除。
- **插入**（insertion）——一个或多个核苷酸插进序列中。

突变也可根据它们是否涉及单一的 DNA 序列（简单突变，节 11.2）或者是否涉及两条等位或非等位序列的交换（节 11.3）来分类。上面所列的三类突变均可由简单突变或序列交换造成。

新突变可发生于单一个体的体细胞或种系中（节 3.1.3）。如果一个种系突变并不严重损害一个人生育能传递该突变的后代的能力，它就能够传播至一个（有性）群体的其他成员。如果某个基因座在人群中存在一种以上频率大于 0.01（高于单纯靠反复突变所能维持的水平）的变异（等位基因），则该等位基因的序列变异通常称作 DNA 多态性。

然而，存在范围自单一核苷酸改变到大范围改变的各种不同类型的多态性（框 11.1）。

框 11.1 遗传多态性与序列变异的分类

个体表型的差异在很大程度上归因于遗传变异，多种不同类型的遗传多态性和大范围的序列变异已经为人们所知。然而，对于表型的影响最终将体现在蛋白质或 RNA 水平上。编码多肽的 DNA 的改变能导致氨基酸的改变，从而产生**蛋白质多态性**（protein polymorphism）。除了 DNA 和已经被研究了多年的蛋白质多态性，还有了解得很少的转录水平上的表达数量差异。等位人类基因的**表达差异**（expression variation）可能缘于包括调节转录及剪接序列在内的调节 DNA 的改变。这类序列变异可能经常是常见病易感性的基础，然而探索人类基因表达中的等位基因变异的定量方法直到最近才开发出来（Yan *et al.*,2002）。常见的 DNA **多态性**（DNA polymorphism）与大范围的序列变异类型介绍如下：

单核苷酸多态性（single nucleotide polymorphism，SNP）

顾名思义，它仅涉及单一核苷酸。对大多数 SNP 来说，单一的核苷酸被一个不同的核苷酸所替换（**核苷酸替换**，nucleotide substitution），但该术语也涵盖了涉及核苷酸插入（insertion）或缺失（deletion）的改变（simple indel polymorphism）。典型的 SNP 仅具有两个等位基因。一些 SNP 将引起限制性酶切位点的改变（**限制位点多态性**，restriction site polymorphism，节 7.1.3）。由于编码 DNA 仅占约 1.5%的人类基因组，大多数 SNP 均见于非编码 DNA 中，诸如位于内含子和基因间序列中。然而，SNP 在染色体上的分布远非均匀：含有很少 SNP 的大片染色体区域常发现于含有许多 SNP 的大片区域旁（节 11.2.6）。SNP 的分型见节 7.1.3。

简单的可变数目串联重复（VNTR）**多态性**

VNTR 多态性在传统上指基因座上包含一段简单序列的串联重复的等位基因。它涵盖了两种类型：**微卫星** = SSR（**简单序列重复**，simple sequence repeat）**多态性**，其中的简单序列为一到几个核苷酸长，整排长度在小于 10 到超过 100 之间；**小卫星** DNA **多态性**（minisatellite DNA polymorphism），所涉及的序列排常常跨越几百个核苷酸，且由长度为 9 到数十个核苷酸之间的序列的串联重复构成。VNTR 基因座上常具有多个等位基因，一些高度可变的小卫星呈现出超乎寻常的变异（例如：MS32 基因座的杂合度数值达 0.975）。

这两类多态性在多肽编码 DNA 中非常罕见。例外包括非常罕见的非移码 SSR 多态性以及非常偶然表达的小卫星多态性。例如：位于 1q21 的 MUC1 基因座据知编码一种见于若干上皮组织及体液中，由于小卫星编码的重复的广泛变异所产生的高度多态的糖蛋白（Swallow *et al.*,1987）。此外，这些多态性中的一部分可能距基因很近，从而能影响它们的表达。一个显著的例子就是 INS VNTR，一个位于胰岛素基因翻译起始点上游 596bp 的小卫星多态性，由一段一致序列为 ACAGGGGTGGGGG 的可变数目串联重复构成。不同的等位基因似乎赋予了对 I 型糖尿病不同的易感性，这可能是由于胰岛素基因表达的不同效应所致（节 15.6.4）。

转座子重复多态性

近 45%的人类基因组是由基于转座子的重复所构成的（节 9.5.1）。它们中的绝大多数已不再活跃，但一些 LINE1，Alu 以及基于 LTR 的转座子仍在活跃转座。作为进化中最近发生的转座结果之一，一些基因座具有多态性。例如，Yb9、Yc1 和 Yc2 Alu 亚家族中的许多成员是如此之新近被插入到人类基因组中，以至于在不同人群中所分析的元件中约有 1/3 具有存在或缺失该 Alu 重复的多态性（Roy-Eagle *et al.*,2001）。

大范围的 VNTR 多态性

由于不等交换或不等性姐妹染色单体交换可产生一类大范围 VNTR 多态性，大范围的串联重复

框 11.1 遗传多态性与序列变异的分类（续）

易于发生拷贝数的变化。例子包括在着丝粒处的 α 卫星重复以及各种串联重复性 RNA 基因如 rRNA 基因，以及 17q21-q22 上 RNU2 基因座上的 U2 snRNA（近乎一致的 6.1 kb 单位的 6～30 次以上的重复）等。一些多肽编码基因簇含有易于发生这类多态性的大型串联重复，例如 21 羟化酶/补体 C4 基因簇（节 11.5.3）。

倒位多态性

对于人类基因组常染色质部分的测序揭示了低至中等拷贝数目序列中相关序列之间具有很高（>95%）序列同源性的许多例子。对一些低拷贝数目重复而言，由于进化上最近（灵长类特异的）的**节段性复制**（segmental duplication）（节 12.2.5），极高的序列同源性可延伸达数十至数百 kb。这类序列能使导致易位和大范围缺失、重复以及倒位的不等重组易于发生。尽管缺失和一些重复常常与疾病相关，且通常延伸跨越百万碱基（但仍为亚细胞遗传学水平）的间隔，大规模的倒位多态性仍可能发生，但并不直接促成疾病（节 11.5.5）。

染色体多态性和大范围的序列变异

一些多态性涉及非编码 DNA 的巨大改变，以至于等位基因可通过传统的细胞遗传学分析来分辨。C 显带（用来识别异染色质；框 2.2）常用来显示染色体 9、16 和 Y 等的特定异染色质区的异染色质的大小变化。一个大型的臂间倒位在正常人群中很常见。断裂点明显位于异染色质序列之外的偶然倒位在正常人群中亦可能很常见。

对人类基因组 DNA 而言，**平均杂合度**（mean heterozygosity，也称**平均核苷酸多样性**，average nucleotide diversity）约为 0.08%：即等位基因序列之间平均每 1250 个碱基中有一个不同。该数字最初是通过对少量的个别基因座进行测序研究所得数据进行平均来估算的（Przeworski *et al.*,2000）。不同基因座的杂合度数值变化的确很大，某些基因，尤其是一些 HLA 基因的多态性格外地高（图 12.29），但对于最近可用的全球人类单核苷酸多态性（SNP）数据的分析却证实了 0.08%的平均杂合度数值（Reich *et al.*,2002）。然而，由于突变率相对较低，一个个体中等位基因序列间的绝大多数差异是遗传、而不是由新突变产生的。

突变是进化的原动力，但它们也可能导致致病性（节 11.4，11.5）。它们可以是某种表型异常的直接原因，也可能导致个体对于疾病易感性的增加。通常的低水平突变因而可视为以导致疾病为代价而容许偶然的进化新颖性与造成某个物种一定比例成员死亡之间的平衡。

突变常常起因于 DNA 复制中的拷贝错误。尽管活体内 DNA 复制的忠实性通常极端地高，但错误添加也会以低频率发生，这取决于正确或错误配对碱基的相对自由能。双螺旋几何结构极其轻微的变化将稳定 G-T 碱基对（具有 2 个氢键；注意 RNA 中频繁出现的 G-U 配对，图 1.7B）。为了降低错误添加的差错率，许多 DNA 聚合酶含有一个完整的 3′→5′外切酶，后者可起到**校读作用**（proofreading activity，表 1.2）。当 DNA 合成中插入了一个不正确的碱基时，DNA 合成将无法继续。相反，3′→5′外切酶活性将每次切除 3′羟基末端的一个核苷酸，直到获得一个正确配对的末端，使 DNA 合成得以继续进行。然而，尽管如此，人类基因组的规模还是对 DNA 聚合酶的忠实性提出了极大的要求：人类细胞的每一次分裂均需要由六十亿个核苷酸组成的序列被精确

复制。

DNA 也会遭到细胞内显著的自发性化学伤害，以及因暴露于自然界电离辐射以及活性代谢产物所造成的损伤。因此，为了使突变率降到最低，必须具有有效的 **DNA 修复**（DNA repair）系统来识别和修复 DNA 序列中的众多异常（节 11.6）。此外，在基因表达过程中 mRNA 序列所产生的错误将受到 **RNA 监督**（RNA surveillance）机制的影响，后者将确保除去含有不恰当终止密码子的 mRNA（节 11.4.4）。

11.2　简单突变

11.2.1　由 DNA 复制与修复错误所致的突变很常见

通过暴露于我们的外部环境中的各种诱变剂或者细胞内环境所产生的诱变剂，能够在我们的 DNA 中诱发突变。对于辐射引起的突变而言，例如，Dubrova 等（1996，2002）曾报道，作为切尔诺贝利（Chernobyl）事故放射性散落物严重暴露的后果之一，高度可变的小卫星基因座上的正常种系突变率增加了一倍。然而，到目前为止，正常情况下突变的最大来源为内源性突变，包括在 DNA 复制和修复中出现的自发性错误。

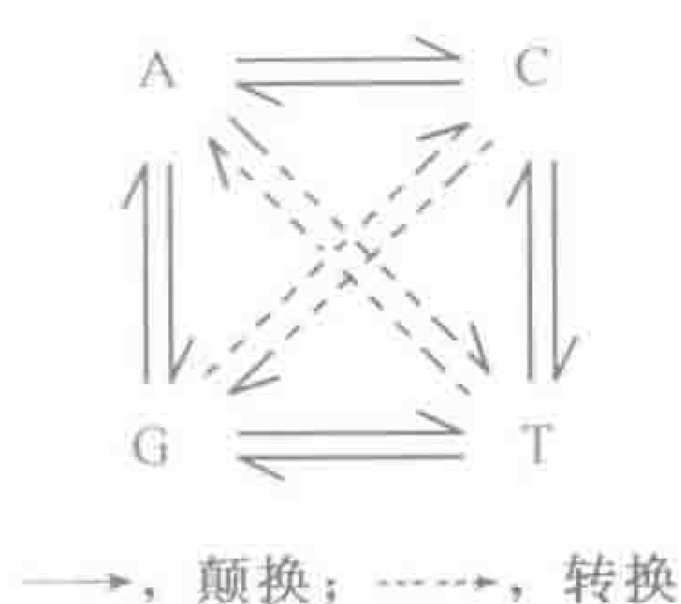
在人类的平均寿命年限中大致将有 10^{17} 次细胞分裂会发生：需要约 2×10^{14} 次细胞分裂来产生成年人体内的近 10^{14} 个细胞，对于某些类型的细胞、尤其是上皮细胞来说，还需要额外的有丝分裂来容许细胞更新（Cairns，1975）。由于每次细胞分裂将需要添加 6×10^{9} 个新的核苷酸，在平均寿命中准确无误的 DNA 复制将需要一个 DNA 复制修复过程，其准确度要足以确保在大约 6×10^{26} 次事件中每次都是正确的核苷酸被插入到生长的 DNA 链中。

要保持如此水平的 DNA 复制保真度是不可能的：的确，观察到的 DNA 聚合酶复制保真度要比这低很多，未经纠正的复制错误以每添加 1 个核苷酸约 $10^{-9}\sim10^{-11}$ 的频率发生（Cooper *et al.*，2000）。由于人类基因的编码 DNA 平均长约 1.65 kb，编码 DNA 突变将以 $1.65\times10^{-6}\sim1.65\times10^{-8}$ 每基因每细胞分裂的频率自动发生。因此，在人类平均寿命中所发生的约 10^{16} 次有丝分裂中，每个基因会是一个约 $10^{8}\sim10^{10}$ 次突变的基因座（但对于任何一个基因来说，仅有极少数细胞会携带突变）。在许多情况下，一个体细胞内的有害基因突变将无关紧要：该突变将可能引起该细胞死亡，却不会对其他细胞产生影响。然而，在某些情况下，突变将可能导致细胞分离的不恰当持续从而导致癌症（第 17 章）。

11.2.2　从替换的类型来看，单个碱基替换的频率是非随机的

碱基替换属于最常见的突变，并可分为两类。**转换**（transition）是指一个嘧啶被另一个嘧啶（C↔T），或者一个嘌呤被另一个嘌呤（A↔G）所替换。**颠换**（transversion）则是指一个嘌呤被一个嘧啶，或者一个嘧啶被一个嘌呤所替换。当某个碱基被另一个所取代时，对颠换来说总是有两种可能，但对转换来说却只有一种可能（图 11.1）。因此，我们可以预测颠换的频率将为转换频率的二倍。

由于在一个群体中等位基因的替换需要几千甚至数百万年来完成，核苷酸替换将无



—→，颠换；----→，转换

图 11.1　颠换频率在理论上预计为转换的二倍

法直接观察到。相反，它们总是通过逐对比较具有共同起源的 DNA 分子，诸如不同物种中的种间同源体来推断。当这完成之后，哺乳动物基因组中的转换率被意外地发现高于颠换率。例如，Collins 和 Jukes (1994) 比较了 337 对人与啮齿类的种间同源体后发现，对于未引起氨基酸改变的替换来说，转换率以 1.4：1 的比例超过了颠换率，而对于确实引起了氨基酸改变的替换来说，该比例则超过了 2：1。

在编码序列中偏好转换甚于颠换是因为它们通常将产生更为保守的多肽序列（下面）。在编码以及非编码的脊椎动物 DNA 中，多于颠换的转换至少在一定程度上可归因于 CpG 双核苷酸中胞嘧啶残基的不稳定性所造成的相对高频率的 C→T 转换。在这类双核苷酸中胞嘧啶常常在第 5 碳原子上发生甲基化，而 5 甲基胞嘧啶则易于脱氨基而成为胸腺嘧啶（框 9.3）。很可能由于这一原因，CpG 双核苷酸是脊椎动物基因组内的一个突变热点：其突变率比普通的双核苷酸高 8.5 倍 (Cooper *et al.*,2000)，而且 CpG→TpG 转换是最常见的致病性点突变。偏好转换甚于颠换的其他因素似乎还包括相关 DNA 聚合酶的序列依赖性校读活性所造成的错配碱基的差异性修复。

11.2.3　编码 DNA 中突变的频率和范围与非编码 DNA 不同

许多突变在本质上是从个体的 DNA 序列中随机产生的。因此，编码与非编码 DNA 序列大致同等地易于发生突变。然而，明确的是，突变的主要后果大部分局限于约占人类基因组 1.5%、据知为编码序列的 DNA，以及占另外 3%左右的高度保守序列（包括调节序列等）。发生于编码序列中的突变可分为两类：

▶ **同义（沉默）突变** [synonymous(silent) mutation] 并不改变基因产物的序列。这仅适用于编码多肽的 DNA。一个同义突变将导致密码子的改变，但由于遗传密码的兼并性，将不引起氨基酸的改变。注意：一些看似沉默的突变可能并非如此，因为它们将影响剪接（通过激活一个潜在的剪接位点或者改变某个外显子剪接增强子序列，节 11.4.3）；

▶ **非同义突变** (nonsynonymous mutation) 将改变基因产物的序列，后者可以是一个多肽或者功能性的非编码（=未翻译的）RNA。

哺乳动物基因组中真正的沉默突变被认为是实质上的**中性突变** (neutral mutation)（并不给所在基因组所属的有机体带来优势或者劣势）。与此相反，非同义突变可以被分为三种：产生有害影响者；无效应者；以及产生有益影响者（例如增进的基因功能或基因-基因之间的相互作用）。大多数新的非同义突变很可能对基因表达具有有害影响，因而可能导致疾病或致命。然而，由于**自然选择** (natural selection，框 11.2)，这类突变在人群中的频率被极大地降低了。在编码 DNA 中的总突变率要比非编码 DNA 低得多，因此，编码 DNA 序列（以及重要的调节序列等）呈现相对较高程度的进化保守。

框11.2　影响群体中等位基因频率的机制

群体中的个体彼此不同，主要是因为所继承的遗传变异。在一个群体中，任何突变等位基因的频率将取决于许多因素，包括自然选择、随机遗传漂变以及非等位基因之间的序列交换等。

自然选择

自然选择（natural selection）是指一些继承下来的遗传变异将造成个体间在顺利存活和繁殖方面能力差异的作用。差异繁殖是缘于个体之间在进行繁殖（受诸如死亡率，健康状况和成功交配等参数的影响）和生育健康后代（受精能力、生育力以及后代生存力方面的差异）能力方面的差别。一种有机体的**适合度**（fitness）是衡量个体存活和成功生育方面能力的指标。在最简单的模型中，一个个体的适合度被认为完全由其遗传组成来决定，并且所有的基因座被想像为各自独立地对个体的适合度起作用，因此每个基因座可以被分开来处理。因此，我们也可以来讨论某个基因型的适合度。

编码DNA中绝大多数的新非同义突变将降低其携带者的适合度。它们因此被选择并且从群体中被淘汰掉（**负性或净化选择**，negative or purifying selection）。偶尔，一个新突变也许与群体中最好的等位基因一样适合，这类突变在选择上为中性。很罕见的情况下，一个新突变将造成选择优势并增加了其携带者的适合度。这类突变将面临**正性**（positive 或者**优势**，advantageous）**选择**，后者可预期将助长它在群体中的播散。如果我们考虑一个具有两个适合度不同的等位基因的基因座，杂合子的适和度将介于两种纯合子之间。在这种情况下，选择的模式将为**共显性**（codominant），而且选择将具有方向性，引起优势等位基因的增加。然而，在某些情况下，一个新突变在纯合子中可能并无优势，而仅在杂合子中具有优势（**杂合子优势**，heterozygote advantage）。这种情况，即杂合子比突变体纯合子和正常纯合子皆具有更高的适合度，属于一种平衡性选择，称为**超显性选择**（overdominant selection）（见框4.8中囊性纤维化的例子）。

随机遗传漂变

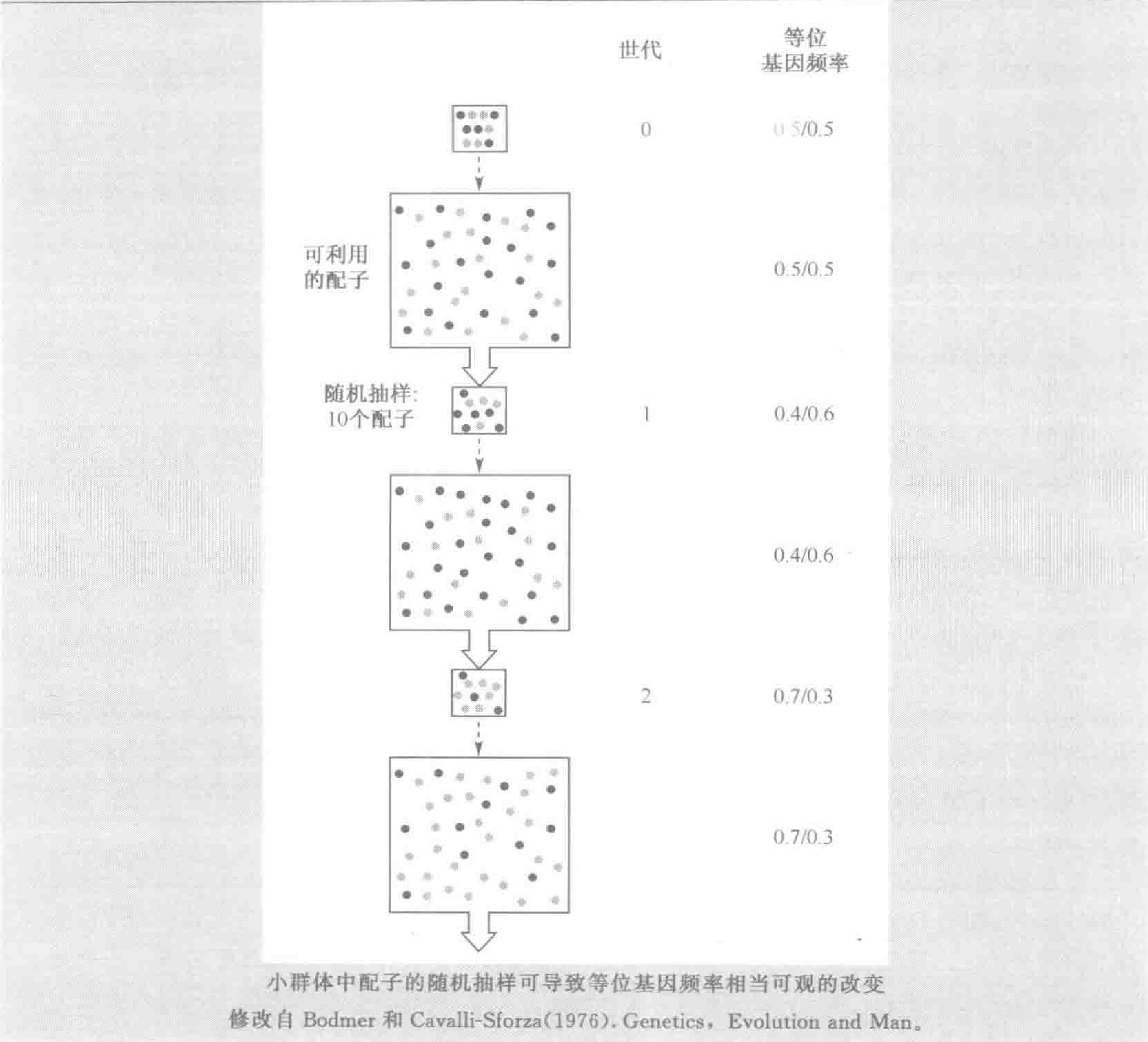
等位基因频率的改变可能完全是随机的（**随机遗传漂变**，random genetic drift）。即使一个群体中的全部个体具有完全一致的适合度，以至于自然选择不起作用，由于配子的随机抽样，等位基因频率仍然会改变。抽样的发生是因为每个世代所产生的配子中仅有一小部分会传递到下一代（由于具体的情况以及选择，并非群体中的所有个体都将生育）。即使没有过剩的配子（因此每个个体将提供两个配子给下一代），抽样仍然发生，这是因为杂合子能产生携带不同等位基因的两种配子，但传给下一代的这两个配子可能碰巧携带了相同的等位基因。

遗传漂变对于大群体没有影响。之所以如此是因为尽管全部配子中只有一小部分被传递下去，但从统计学上看传递的配子已足以在大体上代表群体中的全部配子。然而当群体规模较小时（例如由于地理上的孤立），传递到下一代的配子数量将相应较少，随机遗传漂变就会引起等位基因频率相当可观的改变（图）。在没有新突变和其他诸如选择等影响等位基因频率的因素的情况下，受随机遗传漂变影响的等位基因最终将达到**固定**（fixation，即群体中等位基因频率的数值为0或100%）。

基因座间的序列交换

一些基因家族中的个体基因编码基本相同的产物，然而不同的基因拷贝之间亦可能发生序列交换。例如，人类5.8S rRNA，18S rRNA以及28S rRNA由串联重复的转录单位编码（图10.2），并且尤其容易在不同的重复之间发生序列交换。不同重复之间的序列交换的直接结果之一，就是一种类型的重复在群体中的频率将增高（图11.8是对其一般原理的示意）。在这种多个基因座产生基本相同的产物的情况下，尽管并非孟德尔意义上的等位基因（通常最多只允许在一个二倍体细胞中有两个等位基因），全部的基因在实质上可以被看作是等位基因的等价物。一个特定的重复（“等位基因”）的频率因此能够部分地通过其参与序列交换的频率来确定。

框 11.2　影响群体中等位基因频率的机制（续）



小群体中配子的随机抽样可导致等位基因频率相当可观的改变
修改自 Bodmer 和 Cavalli-Sforza(1976). Genetics，Evolution and Man。

选择压力（selection pressure）（自然选择所施加的约束）将同时减少编码 DNA 中残存突变的总频率以及观察到的突变范围。例如，在非编码 DNA 序列中一个或几个核苷酸的缺失/插入较为频繁，但在编码 DNA 序列中却明显不存在。这是因为这类突变常常将导致翻译读框的移动（**移码突变**，frameshift mutation），导入一个提前终止密码子从而引起基因表达的丢失。即使插入/缺失并未引起移码突变，它们也常常会影响基因功能，例如，通过除去一段重要的编码序列。相反，编码 DNA 的标志则是相对较高频率、发生于对基因表达影响最小的位置上的非随机性碱基替换（见下节）。

11.2.4　编码 DNA 中碱基替换的位置并非随机

发生于非编码 DNA 中的核苷酸替换通常对基因表达并无实际影响。例外则包括启动子元件以及其他一些调节基因表达，或者为剪接所必需的 DNA 序列中的某些改变（图 1.15）。由于保持多肽序列与生物功能的需要，发生于特定多肽的编码 DNA 序列中的替换则呈现一种完全非随机的替换模式。原则上，根据其对编码潜力的影响，碱基替

换可以分为三类（框 11.3）。

框 11.3　多肽编码 DNA 序列中单碱基替换的分类

编码外显子中单核苷酸的替换常常改变 RNA 剪接，导致基因表达的缺陷。这可能以各种方式发生：通过激活外显子中潜在的剪接位点（图 11.12）；影响紧邻保守性 GT 和 AG 信号的剪接供体和受体序列上的核苷酸（图 1.15）；或者改变调节剪接的内部序列，尤其是外显子剪接增强子（节 11.4.3）。其他单碱基替换可以被分为同义（沉默）替换，或者一个密码子发生突变而导致氨基酸序列改变的非同义替换。

同义替换 [synonymous(silent) substitutions]

该替换将产生新的密码子，但仍编码同一个氨基酸。这是编码 DNA 中最为常见的改变（因为它们几乎总是中性突变，并且没有选择压力）。由于第三个碱基的摆动意味着改变了的密码子仍编码跟原来一样的氨基酸，沉默突变大多发生于密码子第三个碱基的位置。但是，偶尔第一个碱基位置也会发生替换，例如在一些亮氨酸密码子（CUA⇔UUA，CUG⇔UUG）以及一些精氨酸密码子（AGA⇔CGA，AGG⇔CGG）中。

无义突变（nonsense mutation）

这些代表了一种形式的非同义替换，其中编码一种氨基酸的密码子被终止密码子所替换。由于这类突变几乎总与基因功能的显著下降有关，选择压力将确保它们在正常情况下很少见。普通的人类多肽由约 500～550 个密码子编码，倘若不存在功能性约束的话，这种长度预计可包含超过 20 个终止密码子。

错义突变（missense mutation）

这些为非同义替换，其中改变的密码子将编码一种不同的氨基酸。它们可以分为两个亚类：

- **保守性替换**（conservative substitution）将导致一种氨基酸被另一种与其在化学上相近的氨基酸所替换。通常，这类替换对蛋白质功能的影响很小，因为新氨基酸的侧链可能与被替换氨基酸的侧链功能相似（见下面的注脚）。为了将核苷酸替换的影响减到最小，遗传密码似乎已进化到了编码相关氨基酸的密码子本身即存在关联。例如，天冬氨酸（GAC，GAT）和谷氨酸（GAA，GAG）的密码子对将确保 GAX 密码子（X 可以是任一核苷酸）第三个碱基的摆动影响甚微。然而，某些密码子第一个位置的改变也可能具有保守性，例如CUX（亮氨酸）⇔GUX（缬氨酸）；
- **非保守性替换**（nonconservative substitution）将导致一种氨基酸被另一种具有不同侧链者所替换（见下表）。有时将造成电荷差异；其他改变可能包括极性侧链被非极性者取代，或者相反的情况。第一和第二密码子位置的碱基替换将常常导致非保守性替换，例如CGX（精氨酸）→GGX（甘氨酸），C CX（脯氨酸），C UX（亮氨酸）或者 C AX（谷氨酰胺/组氨酸）等（X＝任一核苷酸）。

氨基酸对之间的理化差异

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
110	145	74	58	99	124	56	142	155	144	112	89	68	46	121	65	80	135	177	Ser
	102	103	71	112	96	125	97	97	77	180	29	43	86	26	96	54	91	101	Arg
		98	92	96	32	138	5	22	36	198	99	113	153	107	172	138	15	61	Leu
			38	27	68	42	95	114	110	169	77	76	91	103	108	93	87	147	Pro
				58	69	59	89	103	92	149	47	42	65	78	85	65	81	128	Thr

框 11.3　多肽编码 DNA 序列中单碱基替换的分类（续）

续表

Arg	Leu	Pro	Thr	Ala	Val	Gly	Ile	Phe	Tyr	Cys	His	Gln	Asn	Lys	Asp	Glu	Met	Trp	
					64	60	94	113	112	195	86	91	111	106	126	107	84	148	Ala
						109	29	50	55	192	84	96	133	97	152	121	21	88	Val
							135	153	147	159	98	87	80	127	94	98	127	184	Gly
								21	33	198	94	109	149	102	168	134	10	61	Ile
									22	205	10	116	158	102	177	140	28	40	Phe
										194	83	99	143	85	160	122	36	37	Tyr
											174	154	139	202	154	170	196	215	Cys
												24	68	32	81	40	87	115	His
													46	53	61	29	101	130	Gln
														94	23	42	142	174	Asn
															101	56	95	110	Lys
																45	160	181	Asp
																	126	152	Glu
																		67	Met

对于两种氨基酸之间相似性的量化将依据诸如极性、分子体积和化学组成等方面的特性，较大的数值表示更大的差异。在引自 Grantham(1974) 的上表中，最相似的氨基酸对为：Leu⇔Ile(5)，Met⇔Ile(10) 以及 Met⇔Leu(15)；最不相似的氨基酸对为 Cys⇔Trp(215)，Cys⇔Phe(205) 以及 Cys⇔Lys(202)。

框 11.3 所列举的不同类型的碱基替换提示位于密码子第一、第二或第三碱基位置的不同趋势。由于遗传密码子的结构，不同程度的碱基兼并性将是不同位置上的特征。编码氨基酸的密码子上的碱基位置可分为三类：

- **非简并位置**（nondegenerate site）为所有三种可能替换均为非同义性的碱基位置。它们包括除了 8 种以外所有密码子的第一个碱基位置，所有密码子的第二个碱基位置以及两个密码子，即 AUG 和 UGG 的第三个碱基位置（图 11.2）。将已观察到的人类基因中的密码子频率考虑在内，它们占据了人类密码子中 65%的碱基位置。在非简并位置上碱基替换的频率非常低，这与避免氨基酸改变的强大的保守性选择压力一致（见下文）；
- **四重简并位置**（fourfold degenerate site）为所有三种可能替换均为同义性的碱基位置，并见于若干密码子的第三个碱基位置（图 11.2）。它们占据了人类密码子中 16%的碱基位置。四重简并位置上的替换率与内含子及假基因内相似，与同义替换在选择上为中性的假设一致（节 11.2.5）；
- **二重简并位置**（twofold degenerate site）指三种可能替换之一为同义性的碱基位置。它们常见于密码子的第三个碱基位置，但亦存在于 8 种密码子的第一个碱基位置（图 11.2）。它们占据了人类密码子中 19%的碱基位置。与预期一致，二重简并位置

上的替换率居中：三种可能替换中只有一种，即转换可维持相同的氨基酸。其他两种可能替换属于颠换，由于遗传密码子的进化方式，后者常为保守性替换。例如，在谷氨酸密码子 GAA 的第三个碱基位置上，A→G 转换将是沉默的，而两种颠换（A→C；A→T）则将产生一种非常相似的氨基酸，即天冬氨酸的替换。

UU**U**	Phe	17.1	UCU	Ser	14.7	UA**U**	Tyr	12.1	UG**U**	Cys	10.1
UU**C**	Phe	20.4	UCC	Ser	17.5	UA**C**	Tyr	15.5	UG**C**	Cys	12.4
UU**A**	Leu	7.3	UCA	Ser	11.9	(UAA	STOP)		(UGA	STOP)	
UU**G**	Leu	12.7	UCG	Ser	4.5	(UAG	STOP)		UGG	Trp	13.0
CUU	Leu	12.9	CCU	Pro	17.3	CA**U**	His	10.6	CGU	Arg	4.7
CUC	Leu	19.5	CCC	Pro	20.0	CA**C**	His	15.0	CGC	Arg	10.8
CUA	Leu	7.0	CCA	Pro	16.7	CA**A**	Gln	11.9	**C**GA	Arg	6.3
CUG	Leu	40.1	CCG	Pro	7.0	CA**G**	Gln	34.4	**C**GG	Arg	11.8
AU**U**	Ile	15.8	ACU	Thr	12.9	AA**U**	Asn	16.7	AG**U**	Ser	12.0
AU**C**	Ile	21.3	ACC	Thr	19.1	AA**C**	Asn	19.3	AG**C**	Ser	19.4
AU**A**	Ile	7.2	ACA	Thr	14.9	AA**A**	Lys	24.0	**A**G**A**	Arg	11.7
AUG	Met	22.3	ACG	Thr	6.2	AA**G**	Lys	32.5	**A**G**G**	Arg	11.6
GUU	Val	10.9	GCU	Ala	18.6	GA**U**	Asp	22.1	GGU	Gly	10.8
GUC	Val	14.6	GCC	Ala	28.4	GA**C**	Asp	25.7	GGC	Gly	22.6
GUA	Val	7.0	GCA	Ala	16.0	GA**A**	Glu	29.0	GGA	Gly	16.4
GUG	Val	28.7	GCG	Ala	7.6	GA**G**	Glu	40.3	GGG	Gly	16.4

图解
N 非简并位点
N 二重简并位点
N 四重简并位点

图 11.2　人类基因中密码子的频率与非简并、二重简并以及四重简并位置的分布

观察到的密码子频率以千分之几的数值形式给出（例如，UUU＝17.1/1000 或 0.0171）。它们得自密码子使用数据库（http://www.kazusa.or.jp/codon/），涉及对来自 GenBank Release 131.0（2002 年 8 月 15 日）的 21,930，294 个密码子的抽样。**注意**：尽管 61 个第一碱基位置中有 8 个为二重简并，第一碱基位置上所有可能的替换中约 96％为非同义性。在第二碱基位置上的所有替换中，100％为非同义性而第三碱基位置上约为 33％。

表 11.1　相对的氨基酸可变性

（Ala＝100；数据引自 Collins and Jukes，1994）

Thr	116	Asp	84
Ser	114	Lys	77
His	107	Glu	76
Asn	107	Pro	67
Met	102	Leu	58
Ala	100	Gly	57
Gln	99	Phe	55
Val	98	Tyr	53
Arg	94	Trp	31
Ile	92	Cys	29

遗传密码的结构以及一种氨基酸在功能上与另一种的相似程度将影响有关的**氨基酸可变性**（amino acid mutability）。特定的氨基酸可能扮演其他氨基酸所无法轻易取代的角色。例如，半胱氨酸经常参与在形成一个多肽的构象中起重要作用的二硫键的形成（图 1.25）。由于其他氨基酸都不具有含巯基的侧链，在许多位置均存在保留半胱氨酸残基的强大选择压力，并且半胱氨酸也属于可变性最小的氨基酸类（表 11.1）。相比之下，其他特定的氨基酸诸如丝氨酸和苏氨酸具有非常相似的侧链，替换其密码子的第一个碱基位置（ACX→UCX；X＝任意核苷酸）和第二个碱基位置（A CPy→A GPy；Py＝嘧啶）均可引起丝氨酸→苏氨酸的替换。可能由于这一原因，丝氨酸和苏氨酸属于最具可变性的氨基酸

(表 11.1)。

11.2.5 不同基因以及不同基因组分之间替换率变化相当大

中性替换率

最近提供的人及小鼠基因组的序列草图使得在基因组范围分析序列差异和替换率成为可能（小鼠基因组测序协作组，Mouse Genome Sequencing Consortium，2002）。作为一个参考点，中性替换率通过比对非功能性 DNA 来估测。基于转座子的非编码重复序列已鉴定插入到共同祖先的基因组 DNA 中，并且在人和小鼠的歧化之前被固定下来而做到的。总共 165Mb 的原始重复序列被发现（种间同源序列由比对邻近的非重复 DNA 而确定）。

原始重复序列的总体序列一致性估计为 66.7%。在四重兼并位置（上节），即另一套潜在的无功能序列上的序列一致性为 67%。由这些显著相近的数值可推算出中性替换率为每位点 0.46～0.47 次替换（小鼠基因组测序协作组，2002）。考虑到产生现代人类和小鼠的种系中突变率的差别（节 11.2.6），这被推测反映了人类种系中每位点每年约 2×10^{-9} 的中性替换率。

不同基因组分内的替换率

对 14 000 多对种间同源性人和小鼠基因序列的比较突出了不同基因组分内替换率的差别（小鼠基因组测序协作组，2002；基于一套子数据集的直观表示见图 11.3）。同预期相符，编码区域最为保守（序列一致性达 85%，或者说每核苷酸位置 0.165 次替换），但内含子序列的保守性就要差得多（序列一致性为 68.6%，接近于全基因组序列比较所得的 69.1% 总体序列一致性）。非翻译区呈现中等的保守性（5′非翻译区为 75.9%，3′非翻译区为 74.7%），旁侧的 200 bp 上游区（启动子区域为 73.9%）以及 200bp 下游区（70.9%）亦然。

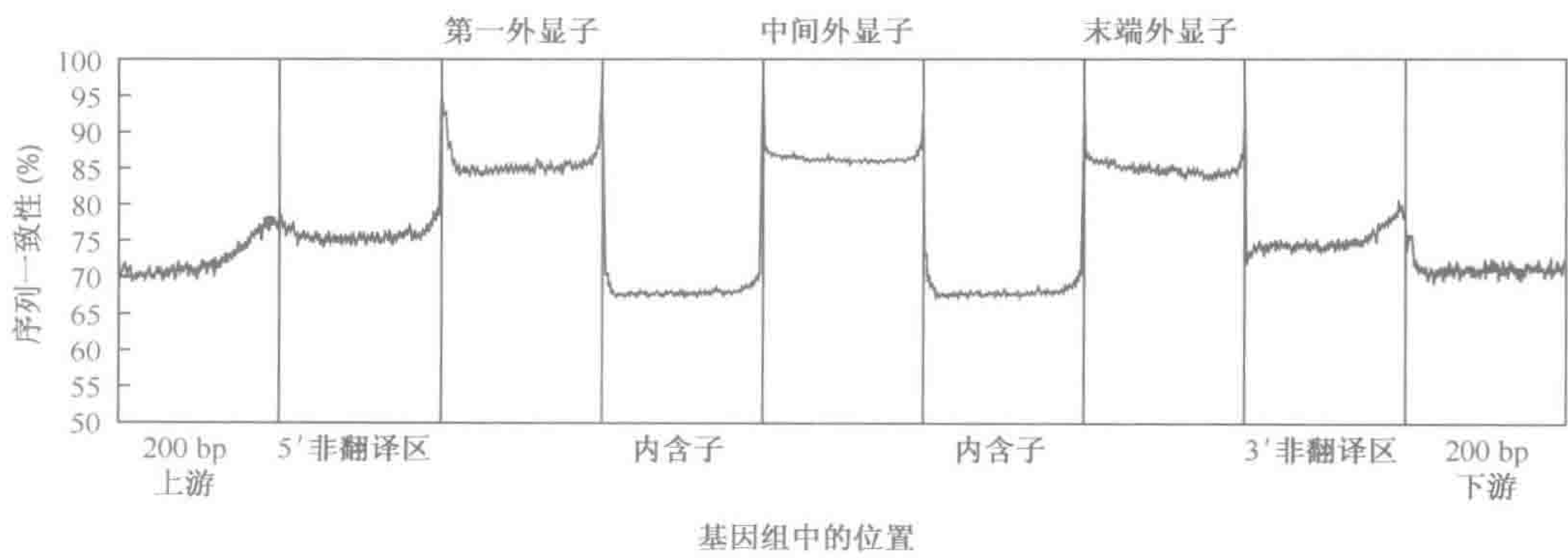


图 11.3 人-小鼠序列保守性在一个典型的基因内的变化

将 3000 多条人类 RefSeq mRNA 及其上下游基因组序列与种间同源性小鼠序列进行比对所显示的核苷酸替换率在一个基因不同组分中的差异。经 Nature Publishing Group 允许，改编自小鼠基因组测序协作组(2002). Nature 420，520～562，图 25A。

不同基因和结构域内的替换率

不同基因的保守程度有所不同。一般说来，同义位点上的替换率相对固定，而非同义位点上的替换率则可能显著变化。如果我们用 K_S表示每个同义位点的替换数，K_A表示每个非同义位点的替换数，受负性（净化）选择的基因将具有 $K_A \ll K_S$。在氨基酸替换方面受正性选择的基因亦可能呈现 $K_A \ll K_S$，因为该正性选择仅对少数至关重要的位点起作用，并被大量的其他位点的净化选择所抵消。小鼠基因组测序协作组在对 12,000 多对种间同源性人与小鼠基因进行检查时，得到的 K_A/K_S中值为 0.115（2002）。

序列极度保守的蛋白质是一个极端，因为它们是细胞关键性功能所必需的，具体的蛋白质包括诸如肌动蛋白、钙调蛋白、组蛋白、核糖体蛋白、泛素等。例如，人类、小鼠以及果蝇的泛素蛋白呈现 100%的序列一致性，而且与酵母的比较亦显示 96.1%的序列一致性。这些基因并未受到特殊的突变防护，因为其同义密码子的替换率与许多蛋白编码基因毫无二致。相反，它们的不同之处在于与其他基因相比极低的非同义密码子替换率（见表 11.2 中的一些例子）。

表 11.2　哺乳动物蛋白质编码基因同义和非同义突变替换率

基因	比较的密码子数目	非同义突变率（$\times 10^9$）	同义突变率（$\times 10^9$）
肌动蛋白 α	376	0.01	2.92
核糖体蛋白 S14	150	0.02	2.16
核糖体蛋白 S17	134	0.06	2.69
醛缩酶 A	363	0.09	2.78
HPRT	217	0.12	1.57
胰岛素	51	0.20	3.03
α 珠蛋白	141	0.56	4.38
β 珠蛋白	146	0.78	2.58
白蛋白	590	0.92	5.16
IgV_H	100	1.10	4.76
生长激素	189	1.34	3.79
$Ig\kappa$	106	2.03	5.56
干扰素 β_1	159	2.38	5.33
干扰素 γ	136	3.06	5.50

源自人-啮齿类动物比较的数据，摘自 Grauer 和 Li(2000)的表 4.1。

呈现最高非同义密码子替换率的基因包括许多与哺乳动物防御及免疫应答系统有关者（表 11.2），以及八个与最高的 K_A/K_S值存在关联的常见蛋白质结构域中的六个（小鼠基因组测序协作组，2002；表 13）。这些例子中的高 K_A/K_S比值提示它们面临减弱的净化选择或增强的正性选择，或者二者皆有。增强的正性选择可能反映了哺乳动物宿主与其病原体之间的一种竞争性的斗争，其中彼此都面临强大的压力来应对另一个基因组内所产生的新事物。在具有最高 K_A/K_S比值的八个结构域中的同样六个被发现于分泌蛋白中，而分泌的结构域具有比细胞核以及细胞质区域高得多的 K_A/K_S比值（表 11.3）。催化结构域似乎具有相对较低的 K_A/K_S比值。

表 11.3 人与小鼠种间同源基因及结构域编码 DNA 的序列保守性和替换率

种间同源区	氨基酸一致率(%)	K_A	K_S	K_A/K_S
全长蛋白质	78.5	0.071	0.602	0.115
含结构域的蛋白质区	93.5	0.032	0.601	0.061
无结构域的蛋白质区	71.1	0.090	0.586	0.155
所有预测的结构域	95.1	0.024	0.627	0.062
催化结构域	96.6	0.015	0.578	0.033
非催化结构域	94.9	0.026	0.635	0.068
核结构域	98.6	0.008	0.655	0.050
分泌结构域	88.9	0.058	0.694	0.091
细胞质结构域	96.7	0.015	0.587	0.041

K_A，非同义替换率；K_S，同义替换率。经 Nature Publishing Group 允许，数据摘自小鼠基因组测序协作组(2002)。Nature 420，520～562 的表 12。

11.2.6 不同染色体区域以及不同世系中的替换率可能不同

不同染色体区域内的替换率和杂合度

出于各种原因线粒体基因组内的替换率远高于核基因组 DNA（节 11.4.2），但后者可呈现相当可观的区域性变化。人和小鼠基因组序列的草图为评价这种情况提供了一个理想的机会。使用对齐的原始重复序列（上节）以及对齐的四重兼并位置，小鼠基因组测序协作组（2002）计算了人类基因组范围内大约 2500 个重叠的 5 Mb 窗口在表面上的中性替换率。对不同染色体上的平均替换率的比较提示区域性的变化很明显。X 染色体上的替换率最低，而常染色体之间则存在显著差异（图 11.4A）。由于 X 染色体花费其三分之二的时间在女性中，低替换率可能反映了男性和女性生殖细胞在分裂次数上的不同（框 11.4）。

替换率（也包括缺失和插入的频率）的亚染色体变化也非常明显，例如在人的第 22 号染色体上（图 11.4B）。这在一定程度上反映了碱基的组成：替换率在 G+C 含量极高或极低的区域看起来确实比较高，但其中的关系较为复杂。以人类基因组 deCode Genetics 高分辨率重组图（Kong *et al.*,2002）为参考，在替换率与重组率，还有单核苷酸多态性 SNP 密度之间发现也存在对应（图 11.4C）。人类基因组中平均杂合度大致为 1250 碱基分之一（Reich *et al.*,2002），而 SNP 密度在基因组不同组分中的变化极大，当各窗口平均为 200 kb 时，杂合率将呈现多至 10 倍的差异（国际 SNP 工作组，2001）。在更高的分辨率下，跨越数以万计碱基对的人类基因组相当大的区域看起来具有固有的高及低序列变异率。对于这些区域，仅一小部分（多至 25%）变异似乎是由局部突变率产生；最大的贡献者为共有的家族史（Reich *et al.*,2002）。

框 11.4 突变率的性别差异以及由男性推动进化的质疑

自从 Haldane 首次观察到导致血友病的大多数突变均产生于男性的种系中，人类突变常被认为偏向于继承自父系，造成了**男性推动进化**（male-driven evolution）的概念（Li *et al.*,2002）。两种主要

框 11.4　突变率的性别差异以及由男性推动进化的质疑（续）

的方法被用来评估男性和女性种系中的相对突变率：分子进化法和直接观察引起疾病的突变的方法：

分子进化法：

通常这将涉及将同源性 X 连锁及 Y 连锁基因与另一物种中的种间同源体进行比较来估算 X 染色体基因（K_{SX}）和 Y 染色体基因（K_{SY}）的同义突变率。与常染色体不同，性染色体在两种性别中将花费不等的时间。位于**假常染色体区**（pseudoautosomal region）之外的 Y 染色体序列将所有的时间都花费在男性中；X 染色体序列则平均花费 2/3 的时间在女性中，1/3 在男性中（女性有两条 X 染色体；男性有一条）。如果用 α 来代表男性突变率对女性突变率的比率，这将等同于来确定相对于女性突变率为 1 时的男性突变率 α。对于大多数 Y 染色体序列而言，突变率因此将为 α。对于 X 染色体序列，将为 $2/3\times1$（女性突变率）加上 $1/3\times\alpha$（男性突变率）$=2/3+\alpha/3$。因此，观察到的 K_{SX}/K_{SY} 比率＝（$2/3+\alpha/3$）/α，因而能被用于估算 α。这将通常造成在涉及灵长类的比较中约为 4～6 的 α 估计值（Li *et al.*,2002；Makova and Li，2002），尽管一些研究得出了更接近于 2 的数值，该数值与生殖细胞分裂次数较少的啮齿类中的同等比较没有太大区别。因此，可能突变并不像以前认为的那样严重依赖于复制错误。

直接观察引起疾病的突变

很明显，此处所估算的突变为一个特殊的子集。分析的样本来自于一个带有新突变的患者及其双亲（传递有缺陷染色体的双亲通过鉴定与疾病基因紧密连锁的标记来鉴定）。在大多数情况下，突变理应通过种系传递，但如果仅有一个血样被鉴定，突变就可能包括一些**受精卵后突变**（postzygotic mutation，不存在精子或卵子中，但产生于合子形成过程中）。

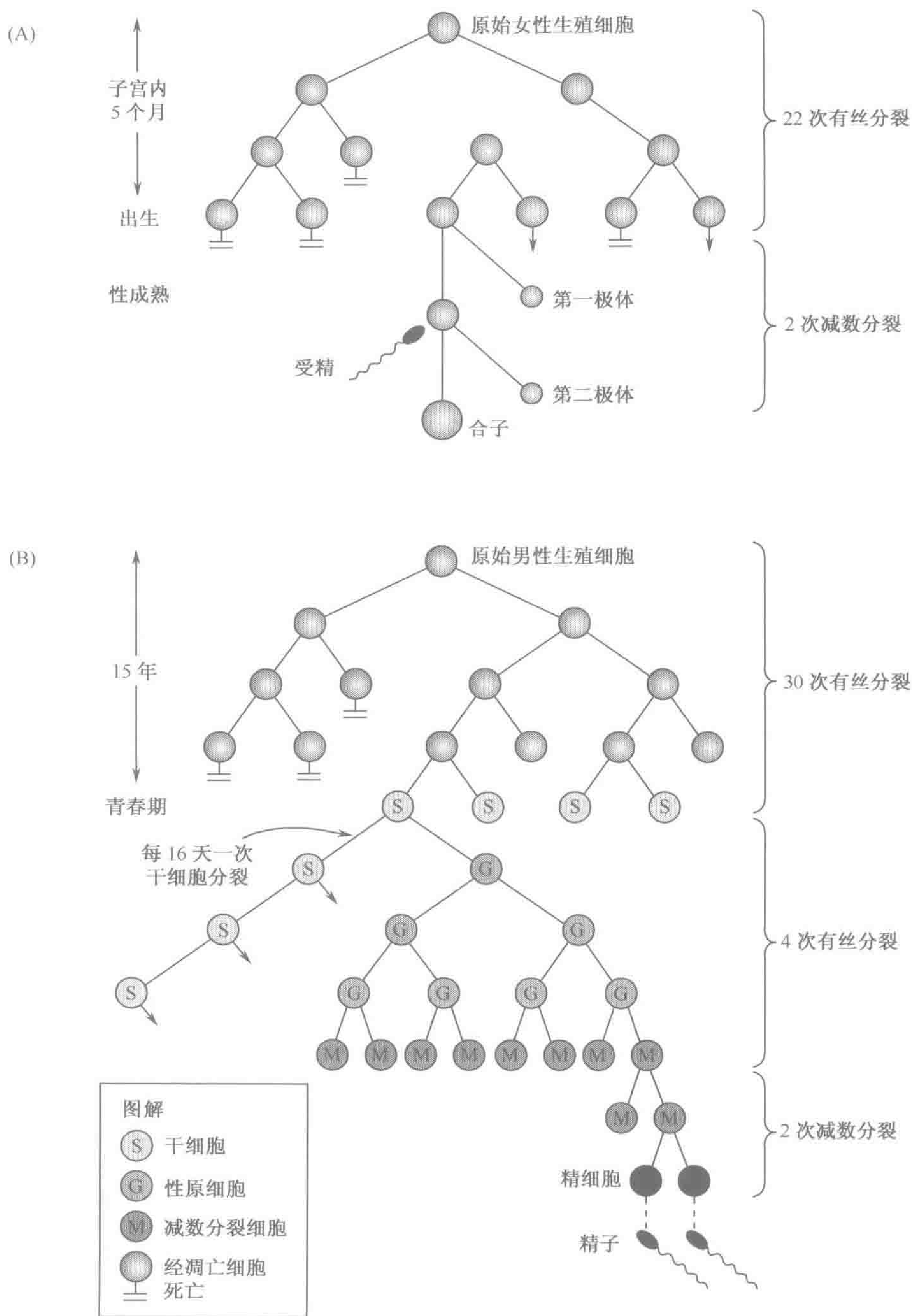
现有数据指向一个普遍倾向于父源突变的明显偏好，至少在单纯点突变的方面（Crow，2000；Li *et al.*,2002 以及表 11.5）。然而，在突变率上的最突出的性别偏倚并不常见，因为它们将涉及编码属于同一蛋白质家族的蛋白质基因的功能获得性突变（FGFR2，FGFR3 和 RET）。此外，一些类型的突变并不显示这样的父源偏倚。例如，抗肌萎缩蛋白基因中极大多数新发生的大范围缺失似乎产生于卵子发生过程中（Grimm *et al.*,1994；见下文），而包含神经纤维瘤蛋白基因的大型缺失常发生于卵子发生过程中（Lopez Correa *et al.*,2000）。

突变率的性别差异可能具有不同的原因（Hurst and Ellegren，1998；Li *et al.*,2002）。在**减数分裂前突变**（premeiotic mutation）（发生于减数分裂前一次生殖细胞分裂中的遗传性突变）中，一个主要的作用因素很可能是人类生殖细胞分裂在次数上的巨大差异。在女性中，从合子到受精的卵细胞的细胞分裂次数恒定不变，因为至发育第 5 个月时所有的卵细胞均已形成，仅需要再进行两次细胞分裂即可产生合子（图，A 板）。从合子到成熟的卵细胞，女性细胞连续分裂的次数被认为约 24 次，这与从合子到青春期生精干细胞估计所需的男性细胞的 30～31 次分裂相近。精子生成接下来还需要进行 6 次细胞分裂，但之后精子发生周期将以每 16 天一轮或每年 23 个周期出现（图，B 板）。正如图中所示，产生精子所需的细胞分裂次数为年龄依赖性。至青春期时，男性生殖细胞将已进行 30 次有丝分裂，精子将在 6 次进一步分裂后产生。之后，精子可连续产生，因为干细胞每 16 天进行一次分裂（＝每年 23 次）。所以如果青春期假定从 15 岁开始，生殖细胞所需的分裂次数在一名年龄为 n 岁的男子中将$=36+[23\times(n-15)]$，在一名 25 岁的男子中约为 265 次分裂，或者在一名 50 岁的男子中约为 840 次分裂。

生殖细胞分裂过程中的 DNA 复制/修复错误预期将产生绝大多数的简单点突变，因此可能预期男性突变率将会比女性高相当多，并且一种父亲年龄效应将较为显著（例如 Crow，2000）。然而，一些更复杂类型的遗传性突变可能更倾向于发生在减数分裂期（**减数分裂突变**，meiotic mutation）而不是减数分裂之前的多次生殖细胞分裂中。例如，大型缺失可能常通过减数分裂期不等交换发生（节 11.3.2）。对没有 Y 染色体同源体的 X 连锁基因诸如抗肌萎缩蛋白的大范围缺失而言，减

框 11.4 突变率的性别差异以及由男性推动进化的质疑（续）

数分裂期突变的显著偏多将意味着大多数缺失应发生于卵子发生阶段。



生殖细胞分裂中的性别差异

(A) 人类卵子发生仅发生于胎儿期并在出生时中止。细胞分裂的总次数被认为约 24 次。(B) 人类精子发生持续于成年时期，精子来源于通过精原细胞自我更新的干细胞。生殖细胞分裂的次数因而完全为年龄依赖性。修改自 Vogel and Motulsky(1996). Human Genetics. Problems and Approaches, 3 rd Edn，获 Spring Verlag. ©1996，Spring Verlag 授权。

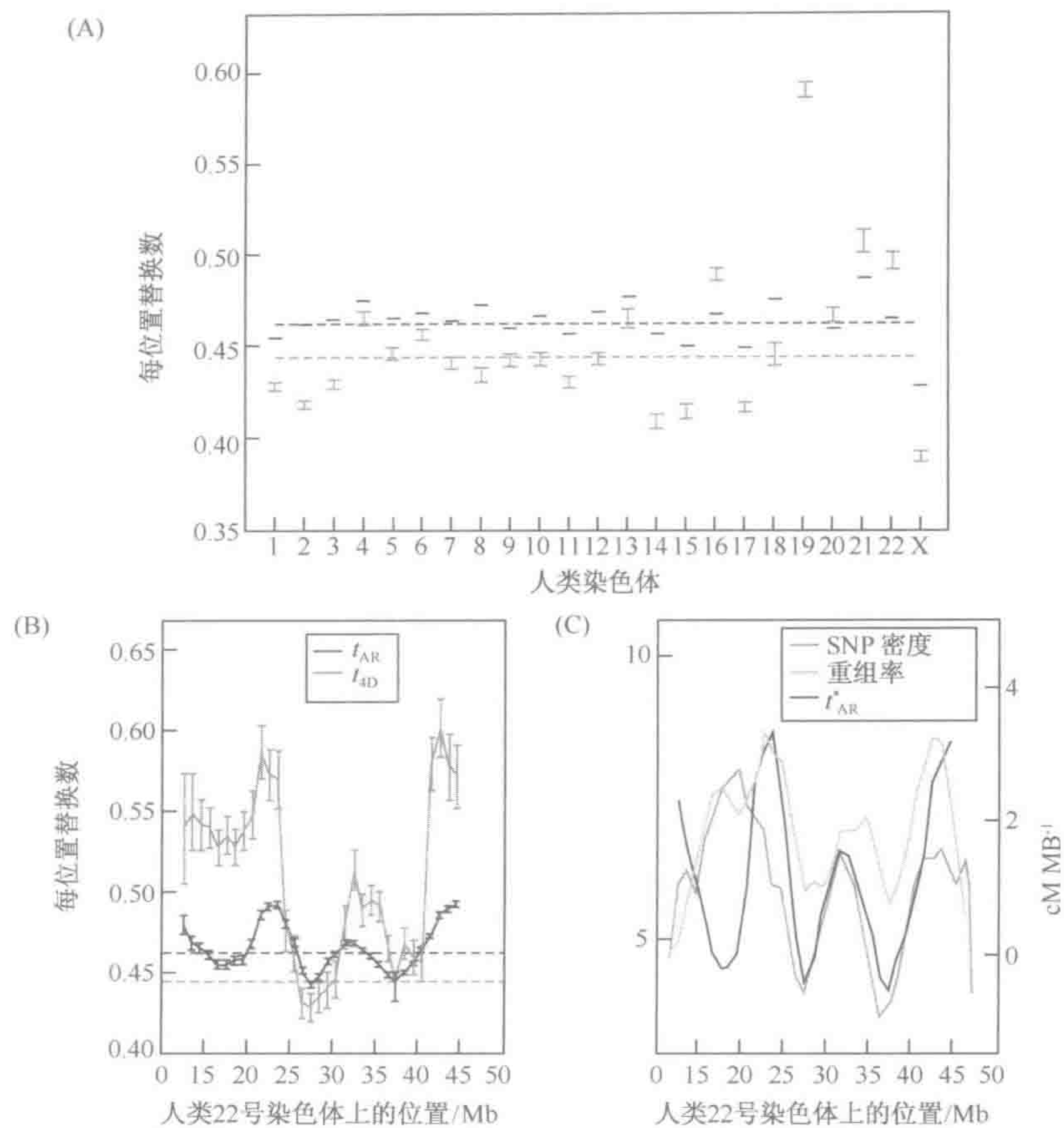


图 11.4　人类染色体与亚染色体区域间替换率的变化

不同人类染色体（A）与人类 22 号染色体（B）上四重简并位置（黑色；t_{4D}）以及原始重复序列位置（深灰色；t_{AR}）内估算的每位置替换数的变化。（A）中虚线代表全基因组的平均值（注意人类 19 号染色体上四重简并位置内显著高的替换数但 X 染色体上低的替换率）(C)22 号染色体范围内原始重复（t_{AR}）内的替换率与 SNP 密度以及重组率的对应关系。经 Nature Publishing Group 授权，改编自小鼠基因组测序协作组（2002）. Nature，420，520～562，图 29A 和 30A。

人与小鼠替换率的比较

人与小鼠的序列草图亦使对于不同种系内替换率如何变化的分析成为可能。由于同义替换从选择约束的角度看来被认为实质上中性，一个恒定的**分子钟**（molecular clock）（借此一个既定的基因或基因产物以恒定的速率进行分子进化）概念在 30 多年前就已提出。然而，从那时起，各种来源的证据曾反对恒定的分子钟（Ayala，1999；以及表 11.2，该表显示同义密码子替换率与非同义密码子替换率一样均存在明显差异）。

为了评估产生现有物种 A 和 B 的两个种系中核苷酸的相对替换率，许多研究均使用了一个亲缘较远的参考物种 C（据知已在进化的更早期、即 A-B 分化之前分支出来）的**相对频率检验**（relative rate test）。对 A 和 C，以及 B 和 C 中种间同源序列的配对比较被用来计算 K 值，即同义替换的频率。K_{AC} 和 K_{BC} 的数值于是将提示产生物种 A 和

物种 B 的种系中相对突变率的水平。这种类型的检验提示产生灵长类的种系中的突变率要低于啮齿类种系，在产生现代人类的种系中就更低了（Wu and Li，1985；Li and Tanimura，1987）。然而，更近的研究（Kumar and Subramanian，2002）对此观点提出了质疑，常见的依据是对于哺乳动物种系发生的理解误差。

为了寻找人和小鼠种系中突变率的潜在差异，小鼠基因组测序协作组（2002）研究了人和小鼠 18 个亚类的原始重复序列的歧化，这些序列在人-小鼠歧化之前不久仍然活跃。各基因组中每个亚类数千个广泛分布的拷贝的比较将估算出自一个一致性序列的歧化，因此替换率的区域性差异能够被最小化。如表 11.4 中代表性例子所示，人的亚类中歧化最小的原始重复呈现与一致性序列 16%的歧化（或者说每位点约 0.17 次替换），而小鼠的亚类则至少发生了 26%～27%的歧化（约为每位点 0.34 次替换）。假设人-小鼠的歧化发生于大约七千五百万年前，人类种系中的平均替换率将为 2.2×10^{-9}，而小鼠种系则为 4.5×10^{-9}。这些数值为人-小鼠歧化以来的平均替换率，而目前小鼠基因组中每年的替换率被认为要高得多（见小鼠基因组测序协作组，2002）。相同的比较亦能够估算出小的（<50bp）插入和缺失的发生率。两个物种均呈现出核苷酸的净损失（缺失碱基与插入碱基的比率在 2∶1 到 3∶1 之间），但在小鼠中总的损失至少要高两倍。

表 11.4　原始重复在小鼠种系中比人类种系中歧化得更快

亚类	替换种类	小鼠		人类		校正比率
		歧化率	替换率	歧化率	替换率	
L1MA6	LINE1	0.28	0.35	0.16	0.184	1.98
L1MA7	LINE1	0.28	0.35	0.16	0.181	1.96
L1MA8	LINE1	0.27	0.34	0.15	0.172	1.96
L1MA9	LINE1	0.28	0.35	0.18	0.201	1.86
MLT1A	MaLR	0.31	0.39	0.21	0.242	1.73
MLT1A0	MaLR	0.30	0.38	0.19	0.219	1.80
MLT1A1	MaLR	0.29	0.37	0.19	0.214	1.78
M1R20	DNA	0.29	0.37	0.19	0.222	1.76
M1R33	DNA	0.27	0.33	0.18	0.211	1.63
Tigger6a	DNA	0.29	0.37	0.18	0.211	1.85

摘自小鼠基因组测序协作组（2002）. Nature，420，520～562。数据显示了 Nature 文章表 6 中 18 个原始重复亚类中具有代表性的 10 个。

表 11.5　导致人类疾病的父源遗传性突变的偏倚

疾病	基因	父源突变数目	母源突变数目	父源/母源突变的比率(α)
X 连锁显性				
Pelizaeus-Merzbacher 病	*PLP*	4	1	4
Rett 综合征	*MECP2*	27	2	13.5
常染色体显性				
软骨发育不全	*FGFR3*	40	0	Inf.
Apert 综合征	*FGFR2*	57	0	Inf.
Crouzon and Pfeiffer 综合征	*FGFR2*	22	0	Inf.

续表

疾病	基因	父源突变数目	母源突变数目	父源/母源突变的比率(α)
Denys-Drash 综合征	*WT1*	2	0	Inf.
先天无神经节性巨结肠病	*RET*	0	3	0
多发性内分泌瘤 2A	*RET*	10	0	Inf.
多发性内分泌瘤 2B	*RET*	25	0	Inf.
神经纤维瘤 2 型	*NF2*	13	10	1.3
Von Hippel-Lindau 病	*VHL*	4	3	1.3
合计		204	19	10

数据摘自 Li 等 (2002). Curr. Opin. Genet. Dev. 12，650～656，获 Elsevier 授权。

11.3　引起重复间序列交换的遗传机制

除非常频繁的简单突变外，还有几类涉及等位或非等位基因之间序列交换的突变，常涉及重复性序列。例如，串联重复 DNA 易于产生缺失/插入多态性，借此不同的等位基因在串联重复完整拷贝的数目上将有所不同。这类**可变数目串联重复**（variable number tandem repeat，VNTR）**多态性**可能出现在重复单位非常短（微卫星）；中等大小（小卫星）或者较大的情况下。取决于重复单位的大小（见下两节），不同的遗传机制均可引起 VNTR 多态性。此外，散布的重复亦可通过各种不同的遗传机制使缺失/重复易于发生。这些将在疾病突变的背景下进行专门讨论，因此放在节 11.4 里。

11.3.1　复制滑移可引起短串联重复处的 VNTR 多态性（微卫星）

微卫星基因座上的种系突变率各不相同，但每位点每代常常在 10^{-3}～10^{-4} 的范围内（Ellegren，2000)。(CA) / (TG) 微卫星以及四核苷酸标记基因座上新长度的等位基因据知无需旁侧标记的交换即可形成。这就意味着它们并非由不等交换而产生（见下文)。相反，由于新的突变等位基因被发现与最初的亲代等位基因相差仅一个重复单位，解释长度变异最可能的机制是一种由**滑链错配**（slipped strand mispairing）引发的序列信息交换。这将发生于双螺旋的两条互补链之间的正常配对被两条链上的重复发生交错改变时，引起重复之间不正常的配对。尽管滑链错配可设想发生于非复制性 DNA 中，正在复制的 DNA 可能为滑移提供更多的机会，因此该机制常常也被称为**复制滑移**（replication slippage）或者**聚合酶滑移**（polymerase slippage）（图 11.5)。除了串联重复间的错配外，滑移复制被设想可通过非连续性重复间的错配产生的大的缺失和重复，并且认为是 DNA 序列和基因组进化的一种主要机制（Levinson and Gutman，1987；亦见 Dover，1995)。短串联重复的致病潜力相当大（节 11.5.1，11.5.2)。

11.3.2　大的串联重复 DNA 单位容易因不等交换或不等性姐妹染色单体交换易发生插入/缺失

同源重组（homologous recombination）指发生于减数分裂期，或偶尔于有丝分裂期相同或非常相似的 DNA 序列之间的重组（**交叉**，crossover），并通常涉及一对同源

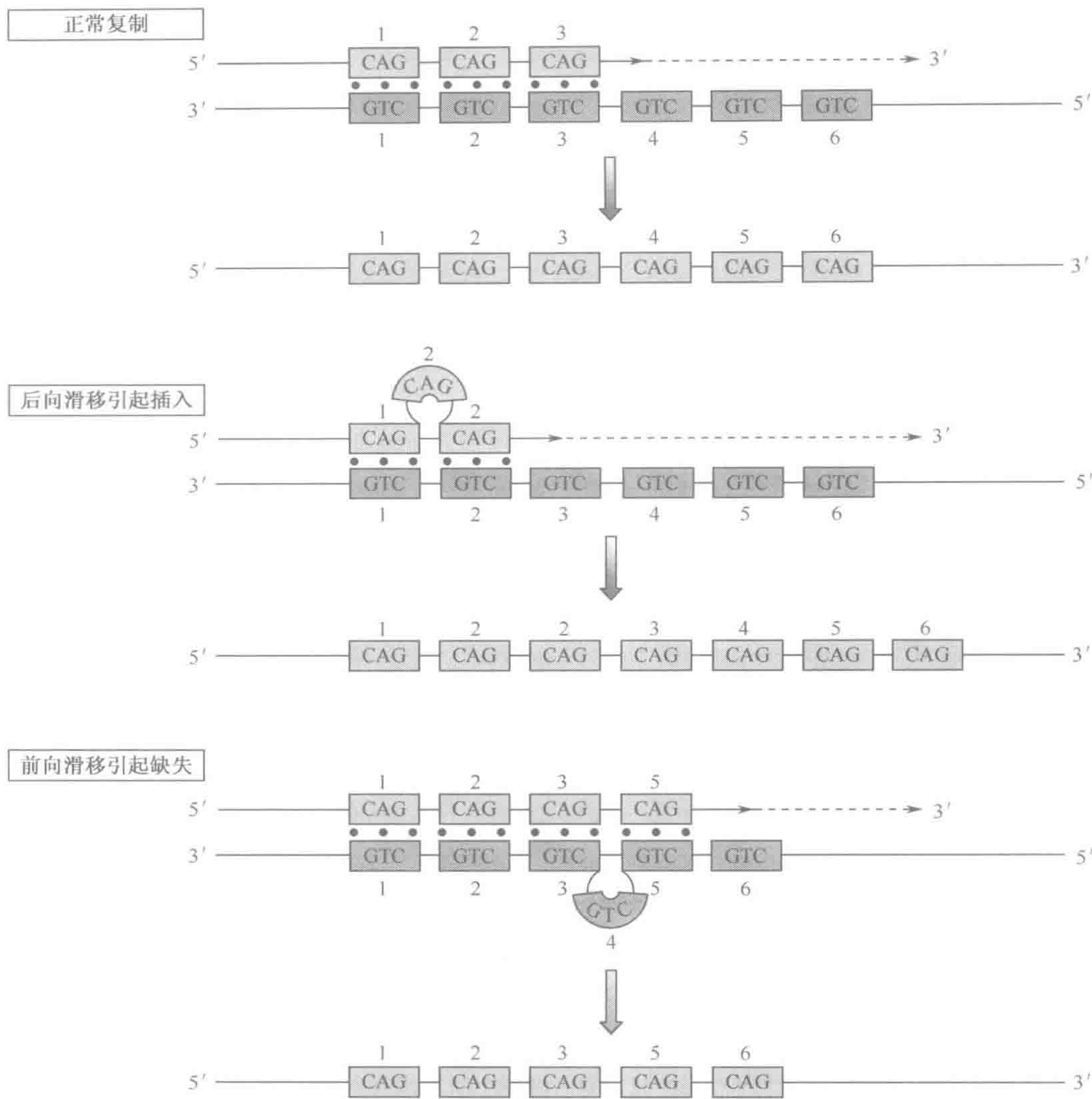


图 11.5　DNA 复制中的滑链错配可引起插入或缺失

短串联重复序列被认为特别容易发生滑链错配（=单一 DNA 双螺旋内互补 DNA 链的错配）。本例显示 DNA 复制中滑链错配将如何发生，下方的链代表亲本 DNA 链，上方的链代表新合成的互补链。此时，滑移将涉及一个非配对区（由空泡表示），含有新合成链（向后滑移）或者亲本链（向前滑移）的一个或更多重复，分别引起新合成链上的插入或缺失。**注**：滑链错配也可能引起非复制 DNA 的插入/缺失。这时将需要两个非配对区，一个含有来自一条 DNA 链的重复而另一个则含有来自其互补链的重复（Levinson and Gutman，1987）。之后一个错配修复酶将通过‘纠正’非配对而引入一个插入或缺失。

基因的非姐妹染色单体的断裂以及片段重接产生的新重组链。**姐妹染色单体交换**（sister chromatid exchange）是一种相似类型的序列交换，涉及单条姐妹染色单体的断裂以及将原本位于相同染色体的不同染色单体上的片段重接。同源重组与姐妹染色单体交换在正常情况下均涉及平等交换——染色单体的分裂与重接发生于各染色单体的相同部位。结果就是，交换将发生于等位基因序列之间并且在等位基因内对应的位置上。在两条等位序列之间发生基因内平等交换的情况下，一个属于**融合基因**（fusion gene，或称**杂种基因**，hybrid gene）的新等位基因将产生，包含一个等位基因的末端片段和另一

个等位基因的剩余序列（图 11.6）。然而，平等的姐妹染色单体交换在正常情况下将无法引起遗传变异，因为姐妹染色单体具有一致的 DNA 序列。

不等交换（unequal crossover，UEC）为一种非等位基因的同源重组，其中交叉将发生于一对同源体的非姐妹染色单体上的非等位序列之间（图 11.7）。发生交叉的序列常呈现非常高的序列同源性，后者被认为能稳定染色体的错配。姐妹染色单体之间的类似交换称为**不等性姐妹染色单体交换**（unequal sister chromatid exchange，UESCE；图 11.7）。

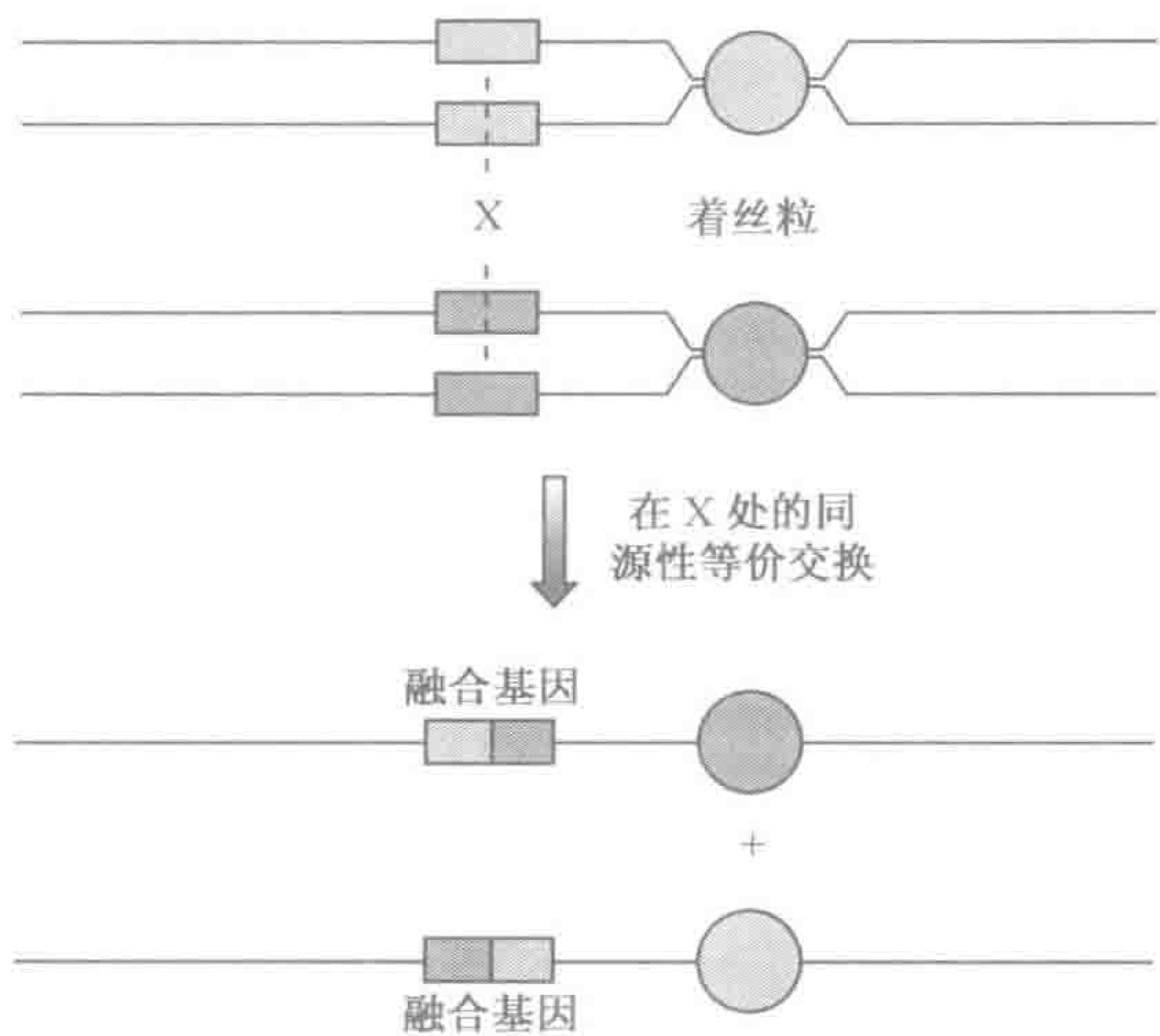


图 11.6 同源等价交换可产生融合基因

本例示意发生于非姐妹染色单体上等位基因之间的基因内等价交换将如何产生由两个等位基因相邻片段构成的融合基因。注意姐妹染色单体上基因之间类似的交换将不会产生遗传新颖性，因为相互作用的姐妹染色单体上的基因序列应该是一致的。

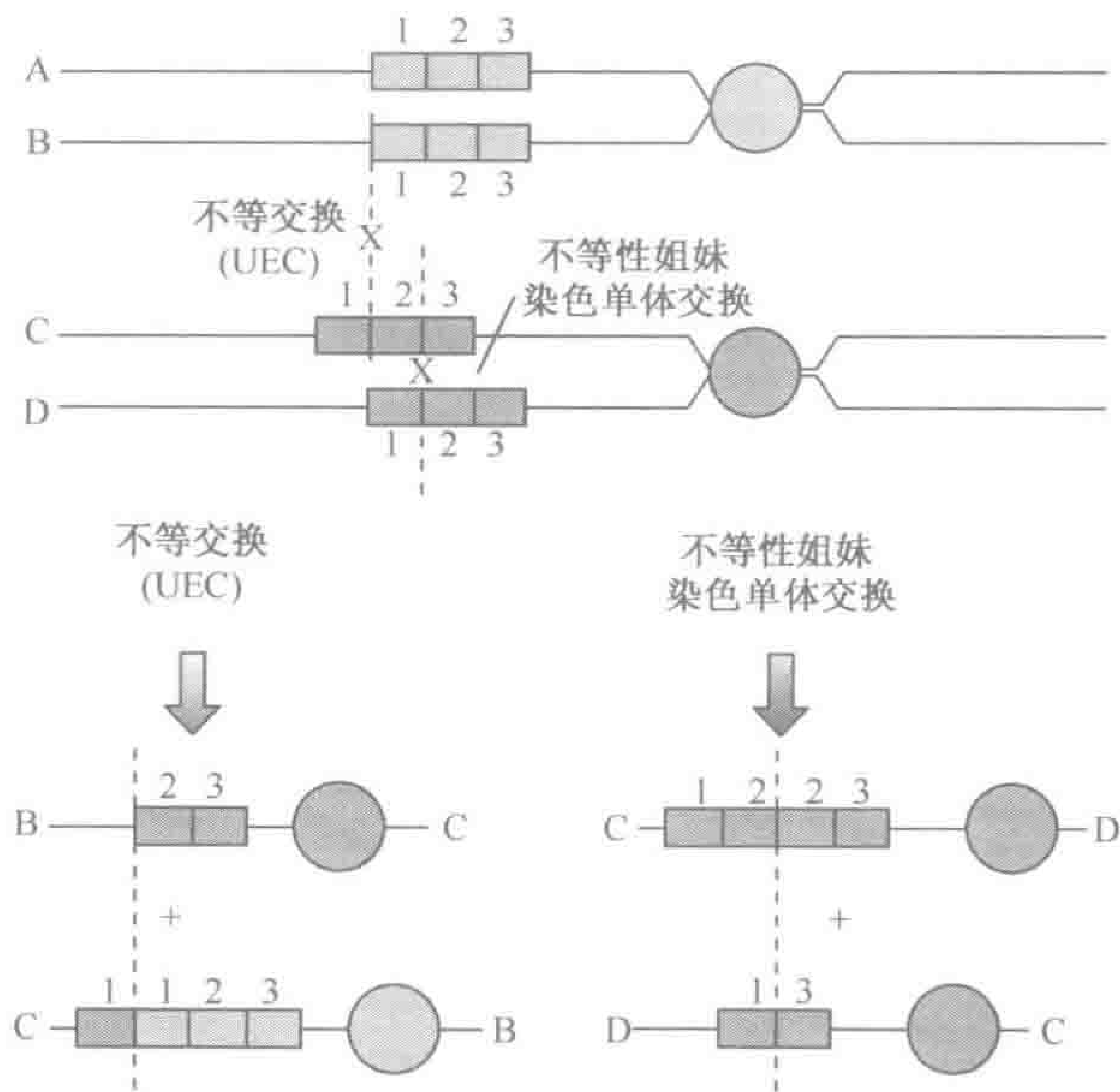


图 11.7 不等交换与不等性姐妹染色单体交换将导致插入和缺失

本例示意了一个串联重复队列中染色单体的不等配对。不等交换将涉及非姐妹染色单体不等配对，随后为染色单体断裂和重接。为了简便起见，染色单体的断裂显示为发生于重复之间，但断裂当然也能发生于重复内部。

注：两种类型的交换均为相互型——一个相关的染色单体将失去一些 DNA，另一个则将获得一些。

UEC 和 UESCE 均主要发生于基因组的一些区域中，后者中存在中等至较长序列的串连重复，重复之间具有高度同源性（例如 rDNA 基因簇、复杂的卫星 DNA 等），在这类情况下，不同重复之间程度极高的序列同源性将促进非姐妹染色单体或姐妹染色单体上非等位重复的异常配对——这些染色单体将与一条染色单体以整倍的重复单位未对齐的形式发生错配。如果在染色单体以这种方式发生错配时发生染色体断裂与重接，那将出现一个相互序列交换，引起一条染色单体上发生插入，另一条上发生同等大小的缺失（可能致病）。这类交换亦可能通过使一段特殊序列经一列串联重复扩展，引起重复单位的均质化，从而导致**共同进化**（concerted evolution）（图 11.8）。

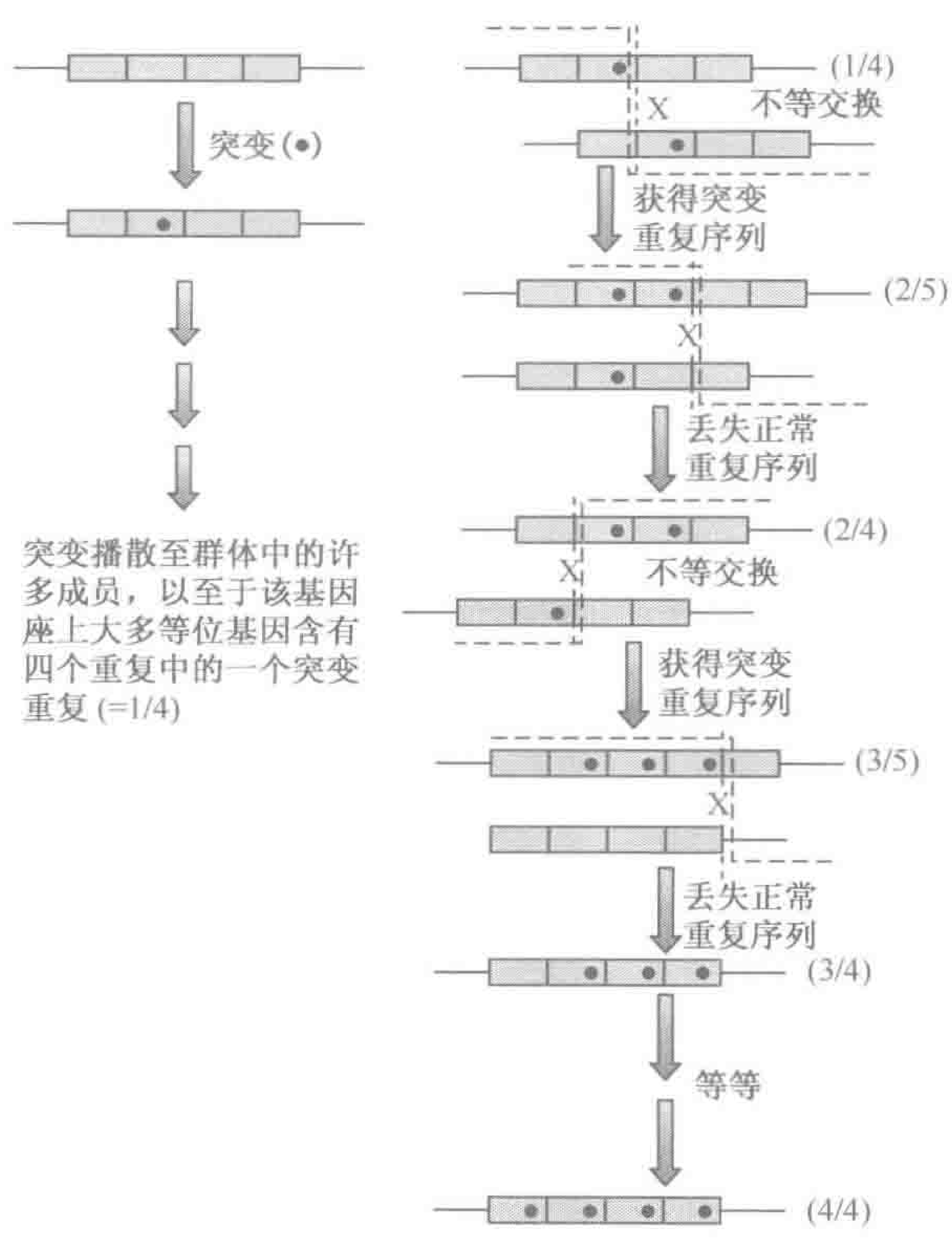


图 11.8　串联重复序列内的不等交换可产生序列均质化

注：一个性别群体中的其他成员的染色体上的相同等位处的新序列变异最初的播散可由随机遗传漂变引起（框 11.2）。一旦突变达到一定群体频率（左图），它就能扩展到该分布区的其他位置（右图）。这可由不等交换（或不等性姐妹染色单体交换）等导致的连续获得突变重复序列引起，有时也可由正常重复序列丢失引起。最终突变重复序列会取代分布区内所有位置的原始重复序列，从而导致突变重复序列的序列均一化。该均一化被认为会导致重复 DNA 序列的种族特异性进化。UEC，不等交换。

虽然 UEC 和 UESCE 在串联重复 DNA 中特别常见，它们也能由相隔相当数量序列的重复的错配引起。例如，非等位 Alu 重复或其他散在重复的错配有时将发生，并可导致一个串连重复基因座形成自最初的单拷贝基因座（图 11.9）。

11.3.3　基因转变事件在串联重复 DNA 中可能相对常见

基因转变指一对非等位 DNA 序列（基因座间基因转变）或者等位序列之间（等位

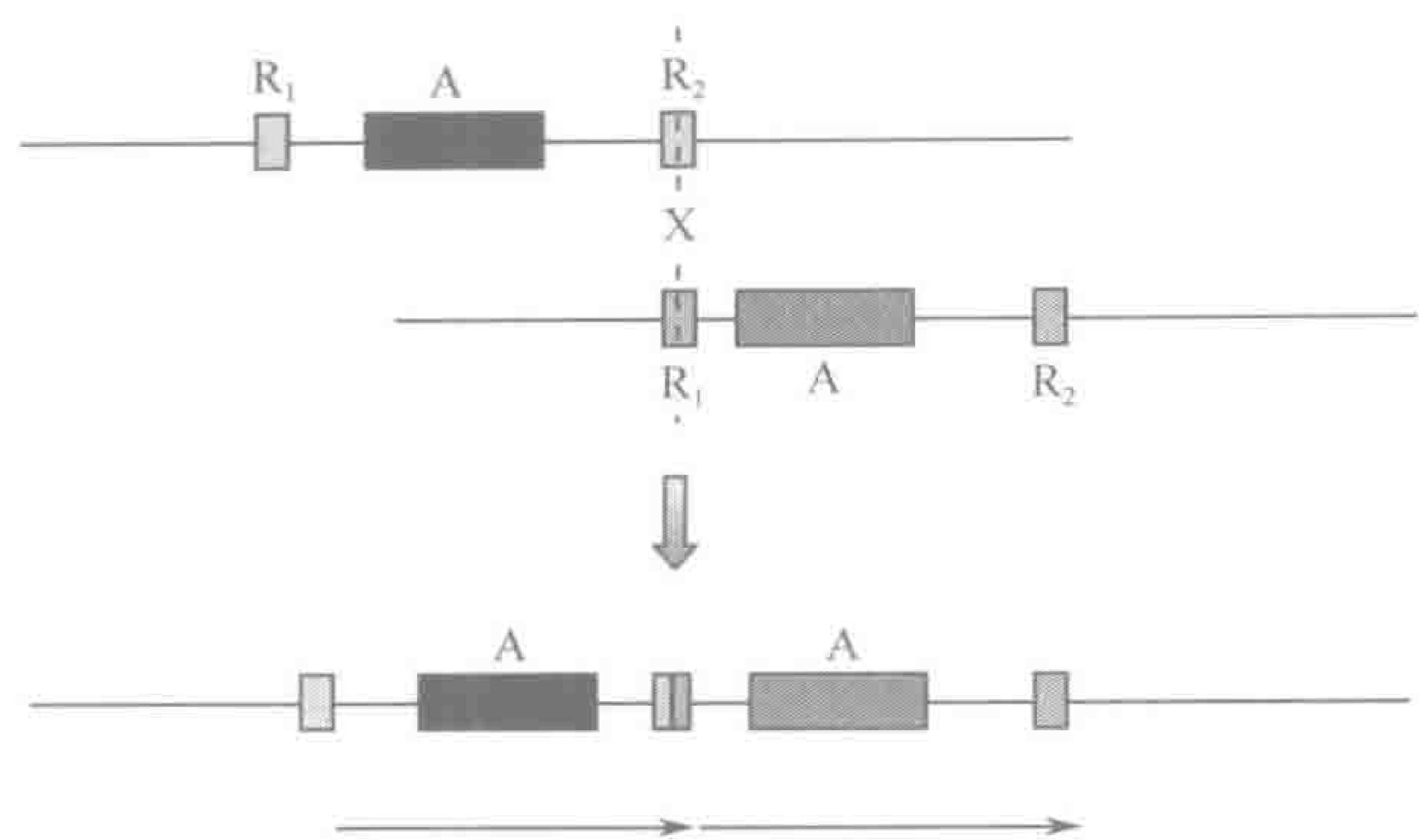


图 11.9　短分散重复序列容易引起不等交换和不等性姐妹染色单体交换，结果导致串联的基因重复
底部的双箭头代表含有一段基因 A 和侧翼序列的串联基因重复的延伸方向。非等位短重复序列（R1，R2）间高度的序列同源性便于染色单体最初的错配。值得注意的是，相同的机制会导致大范围缺失（图 16.3）。

基因内转化）序列信息的非相互性转变。相互作用的一对序列之一，即**供体**（donor）将保持不变。另一条 DNA 序列，即**受体**（acceptor）将因为其一部分或全部序列被来自供体的序列拷贝所替换而改变（图 11.10）。该序列交换因此具有方向性；受体序列被供体序列所修饰，而反过来就不行。

基因转变的一种可能机制被推测为来自供体基因的一段 DNA 链与来自受体基因的一段互补链之间形成了异源双链。异源双链形成之后，受体基因片段的转变可能通过**错配修复**（mismatch repair）——DNA 修复酶将认出异源双链的两条链未完全配对并“纠正”受体链的 DNA 序列使其在转变区内与供体基因链序列完全互补（图 11.10）。

基因转变在真菌中研究得较为透彻，其中减数分裂的全部四种产物均可被回收并研究（四分孢子分析）。在人类和哺乳动物中却不可能这么做。基因转变将无法在高等有机体中被明确地证明，因为它将永远不能与诸如双重交叉等事件相区别（虽然非常靠近的双重交叉一般说来可预期极端不可能）。尽管如此，在哺乳动物基因组中有无数例子，其中一个基因座上的等位基因将呈现一种突变模式，与见于同一物种另一个基因座上的等位基因中者极为类似，提示基因座之间存在基因转变样交换。

尽管对两条序列的简单比较可能会有所启示，但是当一个新突变等位基因可以直接与其始祖序列比较时，基因转变的证据将最为引人注目。某些高度可变的基因座适合于这种类型的分析。特别的是，一些高度可变的微卫星基因座具有较高的种系突变率（通常为每配子 1%或更多），而且单个重复常呈现核苷酸的差异以至于亚类重复能够被识别。种系突变可通过检测和鉴定单个配子中突变的微卫星等位基因来研究。为此，PCR 分析可被进行于对分离自一个人精液的 DNA 的多次稀释所得的等分试样上（**小池 PCR**，small pool PCR），其中每个试样将被校准到含有少数，或许 100 个输入分子。

从个别池中回收的 PCR 产物可被验明以发现任何引起长度可区别于始祖等位基因的新等位基因的新突变。对于三个这类基因座上的种系突变模式的分析未能发现旁侧标记的交换，并提示发生于这些位点的大多数突变具有极性，涉及在串联重复队列的一端优先获得若干重复。存在获得重复的偏倚，证据则是来源于等位基因之间非相互性的序列交换，提示等位基因之间的基因转变（Jeffreys *et al.*,1994）。基因座间基因转变的证

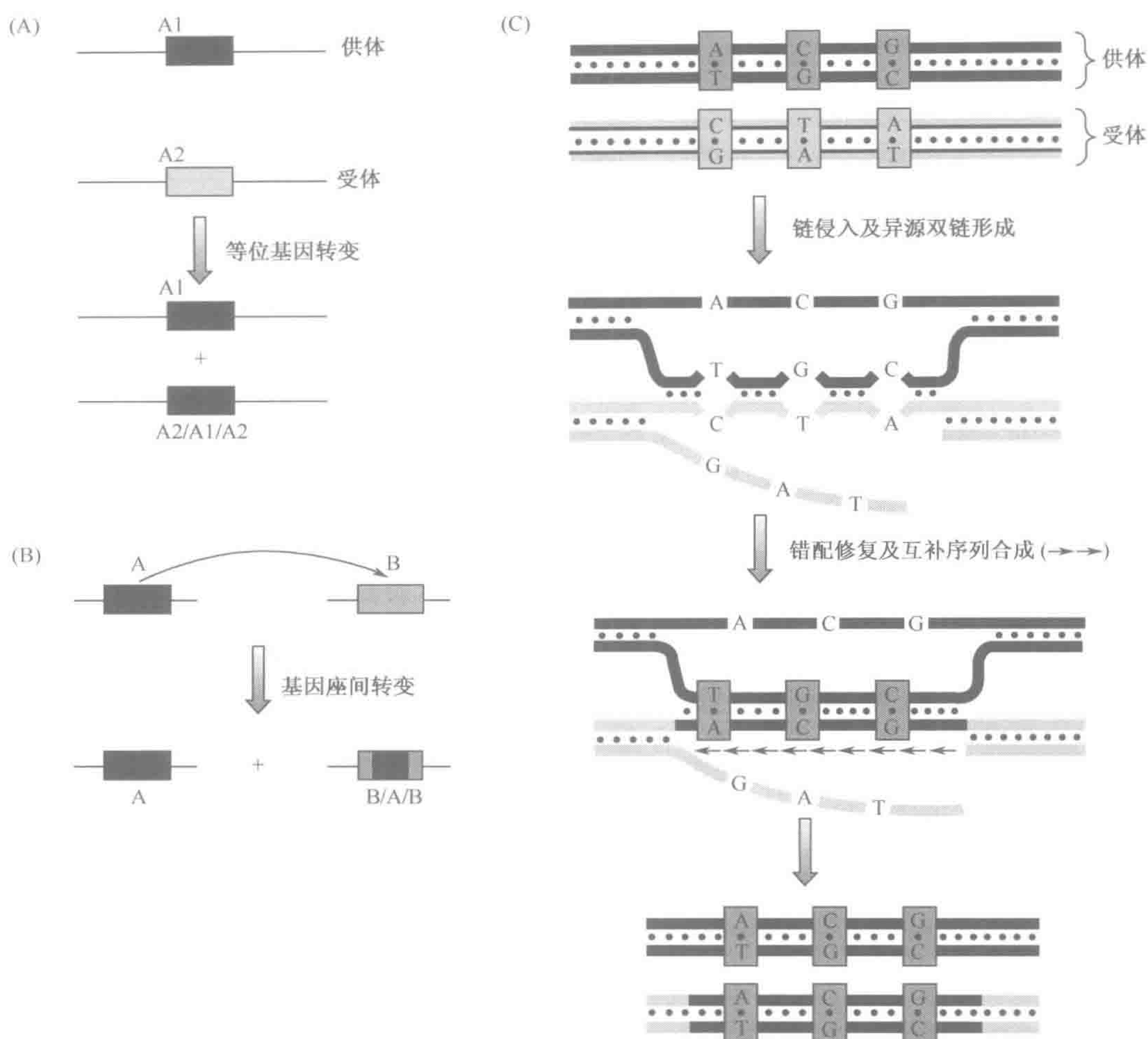


图 11.10　基因转变涉及非相互性序列交换

(A) **等位基因间转变** (interallelic gene conversion)。注意序列交换的非相互性质——供体序列不改变，但受体序列将由于添加了拷贝自供体序列的序列而改变。(B) **基因座间基因转变** (interlocus gene conversion)。这将得力于非等位序列之间高度的序列同源性，正如串联重复的情况。(C) **异源双链错配修复** (mismatch repair of heteroduplex)。这是解释基因转变的几种可能的模式之一。该模式设想供体序列的一条链侵入，并与受体序列的互补链形成异源双链，从而取代受体的另一条链。错配修复酶将识别异源双链内的错配碱基并‘纠正’这些错配，受体序列因而被‘转变’为与供体链序列完全互补。随后的受体链复制和缺口封闭将完成转变。

据亦获得于人类基因中，尤其是类固醇 21 羟化酶基因（节 11.5.3）。

11.4　致病性突变

有害的突变通常将影响基因表达，或者通过直接改变一段编码序列，或者通过改变对基因表达至关重要的基因内或基因外序列（见表 16.1 关于在不同途径中，突变可改变基因表达）。绝大多数被记载的致病性突变均被发现于编码序列中（多半为非同义替

换、无义突变和移码插入/缺失）。由于其相对较高的突变性，CpG 双核苷酸常被定位于编码 DNA 中致病性突变的热点上（Cooper *et al*.,2000）。其他热点包括编码 DNA 内的串联重复（如下文）。在非编码性基因内突变中，剪接位点突变以及非翻译序列中的保守元件内的突变很重要。基因外非编码序列的突变包括启动子及其他调节元件的突变。

11.4.1　原始人类（hominid）中存在较高的有害突变率

中性突变（对于携带的有机体既非有害也非有益）的突变率易于通过先推出一些假定为中性序列的变化的速度来估算（节 11.2.5）。相比之下，**缺失性突变率**（deleterious mutation rate）则更难测定，在 Eyre-Walker 与 Keightley(1999) 所报道的一项研究之前，对于任何脊椎动物并无令人信服的估算。他们研究了 46 种发生于与黑猩猩歧化之后的人类原始种系中的蛋白质。倘若所有的非同义替换均为中性，在他们 46 个基因样本中预期应该有 231 次新替换（假设平均中性突变率为每核苷酸 0.0056 次非同义替换，共研究了 41471 个核苷酸）。相反，仅观察到 143 次非同义替换；其余可预期的 88 次替换推测可能由于有害已被自然选择除去。

Eyre-Walker 和 Keightley(1999) 在存在 60 000 个人类基因的假定之上，估算出一个每人每代 1.6 次突变的有害的速度。按最新估计的 30 000 个基因，编码序列平均长 1.6 kb 重新计算，有害率约占每人每代 2.2 次编码序列突变中的 0.84 次。编码 DNA 占不足 1.5%的人类基因组，按估算的每核苷酸约 2.5×10^{-8}次突变的平均突变率计算，发生于我们的二倍体基因组的总突变数被推算为约每代 175 次（Nachman and Crowell, 2000）。

11.4.2　线粒体基因组是致病性突变的热点

由于人类核基因组很大，大多数突变均发生于核 DNA 序列中。相比之下，线粒体基因组是突变的小靶子（约为核基因组大小的 1/200 000）。与核基因不同，线粒体基因在每个人类体细胞中都有数千份拷贝。某些细胞，诸如脑细胞与骨骼肌细胞具有特别高的氧化磷酸化的需求，因此具有更多的线粒体。卵细胞很特殊，拥有大约 100 000 个 mtDNA 分子，比体细胞多很多。

在正常个体中，～99.9%的 mtDNA 分子是相同的（**同质性**，homoplasmy）。然而，如果产生一个新突变并在线粒体 DNA 群中传播，就将出现两种明显常见的 mtDNA 基因型（**异质性**，heteroplasmy）。假设线粒体 DNA 的突变必须发生在单个 mtDNA 分子上，我们可能将凭直觉预期单一的 mtDNA 突变被保留下来的几率将会很低，而突变率也会相应地低。在此基础上，我们可以预测缘于线粒体基因组内致病性突变的临床疾病的比例应极低。然而，“线粒体疾病”的频率却颇高（节 16.6.6），并且线粒体基因组亦可被视为一个突变热点。这大体上也符合动物线粒体 DNA 的情况，已报道线粒体 DNA 确定的突变率约大于核基因组中同等序列突变的十倍。(Brown *et al*.,1979)。

对于 mtDNA 的高致病力和可变性存在几种解释。93%的线粒体 DNA 为编码 DNA（而核 DNA 仅为 1.6%）。由呼吸链产生的活性氧中间体被认为可引起 mtDNA 大量的氧化损伤（与核 DNA 不同，mtDNA 未被组蛋白保护）。mtDNA 还需要进行较染色体

DNA 多很多轮的复制。线粒体缺乏足够的 DNA 修复机制，尽管若干了解得很清楚的 mtDNA 修复系统目前已知，一些常见的突变，包括胸腺嘧啶二聚体（节 11.6）将无法修复。

对于 mtDNA 突变如何被保留下来的问题需要用发育上的 **mtDNA 瓶颈**（mtDNA bottleneck）来解释。在卵子发生过程中，生殖细胞的数量和每个细胞内线粒体数量具有很大的波动。因此，在三周的女性胎儿中，存在大约 50 个原始生殖细胞，每个细胞各有约 10 个线粒体，但细胞数目的快速增长将随即发生，以至于到第 9 周时，胎儿将拥有超过 50 万个卵原细胞，各具有 200 个线粒体。上述瓶颈被认为发生在介于原始生殖细胞阶段与卵子发生期之间的一个很早的阶段，此时极少数的原始生殖细胞将迁移至性腺，突变 mtDNA 分子可能通过随机漂变和可能通过选择而固定下来（Jenuth *et al*.，1996；Chinnery *et al*.，2000）。

11.4.3 大多数剪接突变将改变正常剪接所需的保守序列，但一些将发生于剪接正常不需要的序列中

许多基因将自然经历不同形式的 RNA 剪接。除此之外，突变有时会引起一种异常的致病性 RNA 剪接形式。有时这将导致完整的外显子序列被排除在成熟 RNA 之外（外显子跳跃，exon skipping；如下）或者保留完整的内含子。在其他时候，异常的剪接模式可能会排除一个正常外显子的一部分或引入新的外显子序列。改变正常情况下为 RNA 剪接所需的保守序列的点突变相对常见。然而，基因的异常剪接有时可由类似剪接供体（=5′剪接位点）或剪接受体（=3′剪接位点）序列但正常不涉及剪接的其他序列元件的突变诱发。

突变改变对剪接起重要作用的序列

对剪接至关重要的保守序列包括：基本恒定不变的 GT 和 AG 双核苷酸，分别位于内含子始端（5′）与末端（3′）；紧邻内含子和外显子序列的剪接供体和剪接受体处序列，包括内含子末端之前的**多嘧啶束**（polypyrimidine tract）；以及**剪接分支位点**（splice branch site）（图 1.15）。此外，剪接据知还受在外显子和内含子中均存在的剪接增强序列（正性调节）和剪接沉默序列（负性调节）的调节。

外显子剪接增强子（exonic splice enhancer，ESE）序列为离散而退化、据知可与剪接调节蛋白结合的长约 6～8 个核苷酸的序列（Blencowe，2000；节 10.3.2）。它们存在于大多数，即使并非全部的外显子中（基本型与可变型剪接均有），以及一组最近被预测存在于人类外显子中的 10 个 ESE 基序，其中一些对调节剪接供体或剪接受体的识别具有重要作用（Fairbrother *et al*.，2002）。对于**外显子剪接沉默子**（exonic splice silencer，ESS）序列以及其他剪接沉默子的了解要少得多（Fairbrother and Chasin，2000）。

改变对剪接起重要作用的序列的突变可导致不同的后果，具体如下：

- **内含子保留**（intron retention）——由于剪接的完全失效所致。这更可能发生于当内含子较小，且邻近序列中缺少可供选择的合理剪接位点或**隐蔽剪接位点**（cryptic splice site，类似于一致性的剪接位点序列相似，但通常未被剪接装置使用的序列，

图 11.11A)。剪接是从含有内含子的基因中有效输出 mRNA 所必需的（Luo and Reed，1999），因此 mRNA 中内含子的保留通常意味着该 mRNA 将被留在细胞核中，以避免与翻译机制相接触（如果被翻译的话，将可能出现不正常的氨基酸或翻译读框的移码)；

▶ **外显子跳跃**（exon skipping）——剪接装置将使用另一个合理剪接位点。剪接供体序列的突变常引起其上游的外显子跳过；而剪接受体序列的突变则常引起其下游的外显子跳过（图 11.11A)。然而，值得注意的是，也可能出现其他结果，包括使用可供选择的外显子或内含子内的隐蔽剪接位点（见下文)，因此结果并非总是容易预测——例子见 Takahara 等（2002)。当一个外显子被跳过时，各种结果都有可能。如果该外显子中的核苷酸数目不能被 3 除，移码将引起一个提前出现的终止密码子，常常导致一个不稳定的 RNA 转录物而不产生多肽。如果外显子跳跃并不引起移码，取决于这些氨基酸对于蛋白质功能和/或结构的重要性，正常编码的氨基酸的缺乏常常将导致一个无功能或异常的多肽。

对 RNA 剪接通常并不重要的序列突变

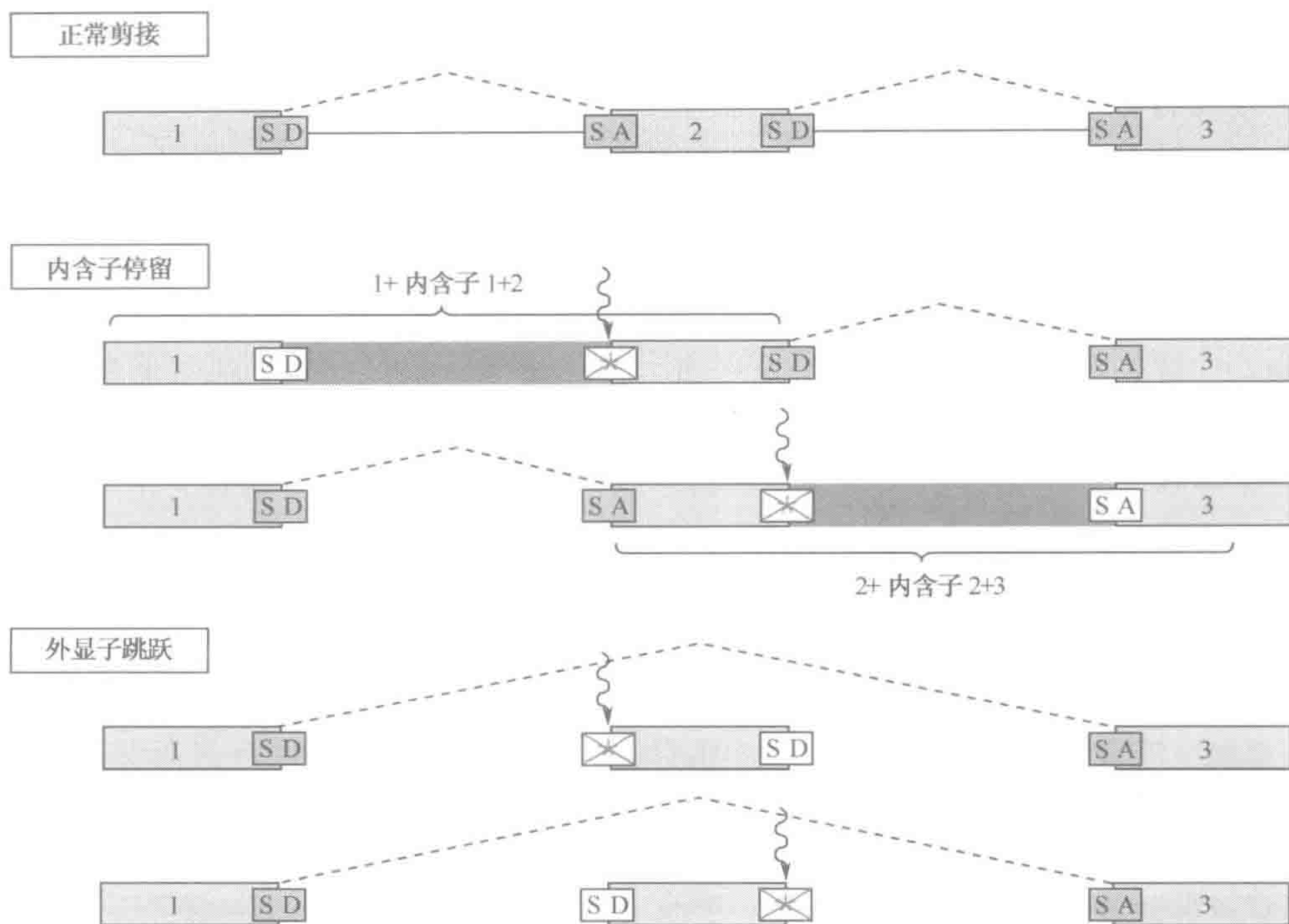
隐蔽（或潜在的）剪接位点［cryptic(latent) splice site］碰巧与真正的剪接位点的序列类似，但一般不用于剪接，除非：（Ⅰ）突变直接改变了序列以至于剪接装置此时将其识别为一个正常的剪接位点（直接激活)；（Ⅱ）真正的剪接位点发生了突变，引起剪接供体或剪接受体的缺陷，在这种情况下，剪接装置将扫描可能的替换并选择一个隐蔽剪接位点（间接激活；见以上关于外显子跳跃的部分)。由于单个剪接供体和剪接受体序列常呈现一些不同于一致性序列的变异，如图 1.15 所示，隐蔽剪接位点频繁出现于基因内部。倘若发生改变的 mRNA 被翻译，内含子中隐蔽剪接位点的启用将引入新的氨基酸；外显子隐蔽剪接位点的启用则将引起编码 DNA 的缺失（图 11.11B)。

外显子中的隐蔽剪接供体以及内含子中隐蔽剪接受体激活的实例分别见图 11.12 和 11.13。前者的警示是外观上的沉默突变仍可能为致病性。值得注意的是，在某些情况下发生于外显子内但不在隐蔽剪接位点上的突变亦能导致该外显子被跳过（见下节)。

11.4.4　引入提前终止密码子的突变常常导致不稳定 mRNA，但也有其他可能的后果

几种不同类型的突变均能引入提前终止密码子（链终止突变)。无义突变仅靠将一个正常密码子用终止密码子来替代就能产生一个提前终止密码子。移码插入和缺失通常也将在突变不远处的下游引入一个提前终止密码子。这是因为没有选择压力来避免在另一个翻译读框中产生终止密码子，因此，以假定确立的核苷酸突变率，在突变位点下游的一段 100 个核苷酸序列中通常会遇到至少一个终止密码子。各种剪接位点突变也会引入提前终止密码子，例如通过跳过一个含有无法被 3 除的若干个核苷酸的外显子。对于链终止突变，基因表达具有几种可能的后果：

▶ **不稳定 mRNA**（unstable mRNA)。这是迄今所知的最常见后果。携带位于最末一个剪接接口的上游至少 50 个核苷酸处的提前终止密码子的 mRNA 在活体中将很快被一种称为**无义突变介导的 mRNA 蜕变**（nonsense-mediated mRNA decay，NMD）的 RNA 监视机制所降解（Lykke-Andersen *et al*.,2001；Maquat，2002)。这将能避免



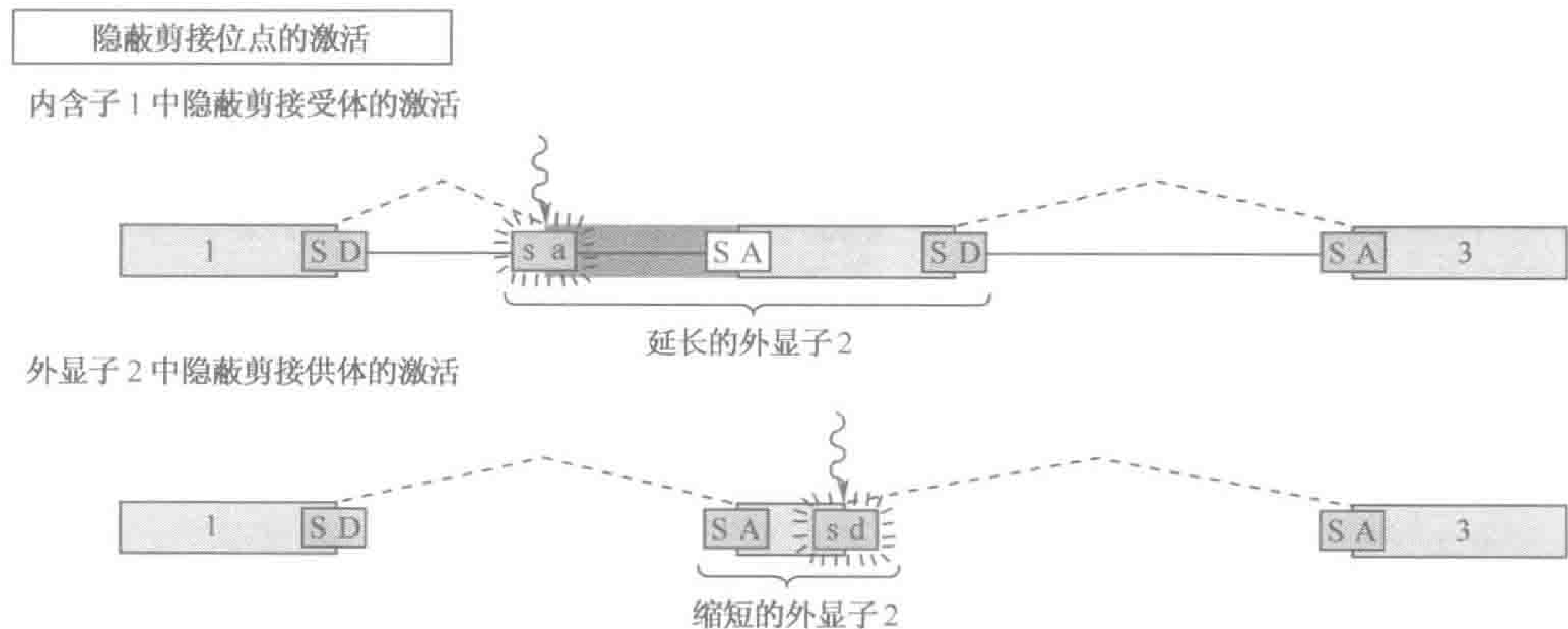


图 11.11　通过保守剪接信号改变或隐蔽剪接位点激活而产生的剪接突变

(A) **保守剪接信号的改变**。剪接供体或剪接受体序列突变（一致性序列见图 11.5）可导致：(a) 剪接失效而介于其间的内含子未被切除所致的内含子保留；(b) 剪接小体将非相邻外显子的剪接供体和剪接受体位点凑在一起所致的外显子跳跃。**注**：优先于启用另一个合理的剪接位点，剪接位点突变偶尔可能造成一个可选择性隐蔽剪接位点（一般未用于剪接）的间接激活（例如 Takahara *et al*.,2002）。(B) **隐蔽剪接位点的直接激活**。通过改变其序列以至于使其与一致性剪接供体或受体序列更相像，突变可直接激活一个隐蔽剪接位点。改变了的隐蔽剪接位点此时可被剪接体识别并使用。外显子与内含子隐蔽剪接位点激活的例子分别见图 11.12 和图 11.13。

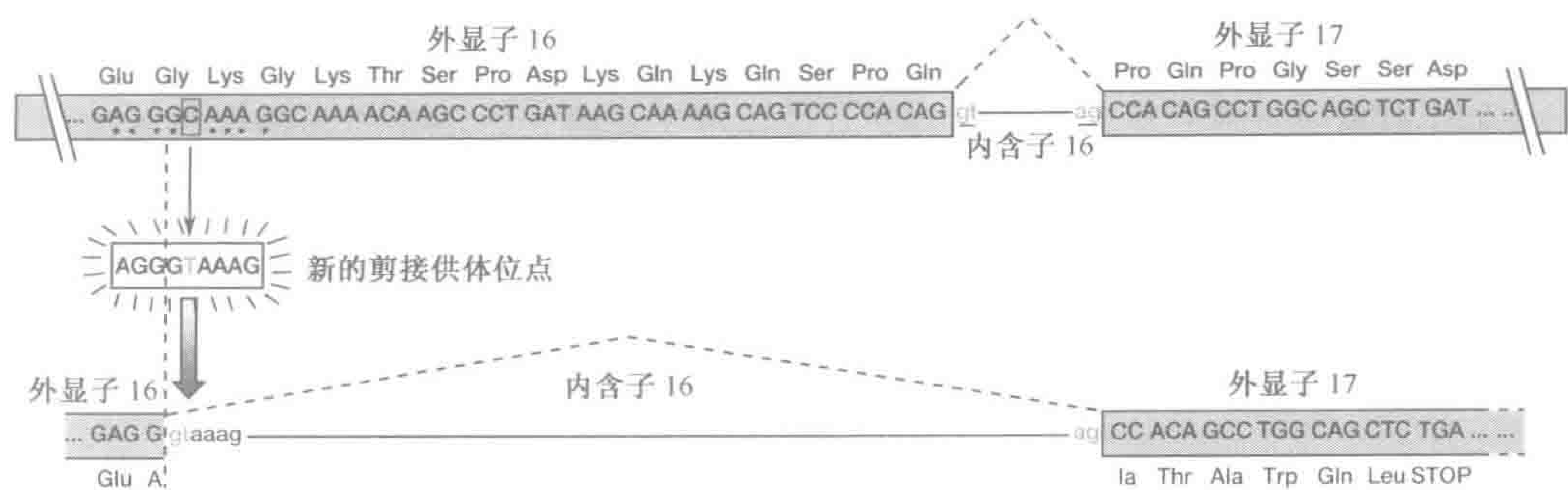


图 11.12 当沉默（同义）突变不再沉默

本例显示了一个发现于一名 LGMD2A 肢带肌营养不良患者中的突变。该突变被发现于这种形式的肌营养不良的已知基因座、钙激活蛋白酶 3 基因内，但发生于一个密码子的第 3 碱基位置，并似乎是一个沉默突变。它应该导致一种甘氨酸密码子（GGC）被另一种甘氨酸密码子（GGT）所取代。然而，该突变仍然被认为具有致病性。这种替换导致了第 16 外显子内的一个隐蔽剪接供体序列（AGGG CAAAAG）的激活，导致第 16 外显子编码序列的丢失和一移码引入的异常剪接。见 Richard and Beckmann(1995)。**注**：致病性同义突变的另一种可能性是通过改变外显子剪接增强子序列而引起它们的效应（节 11.4.3）。

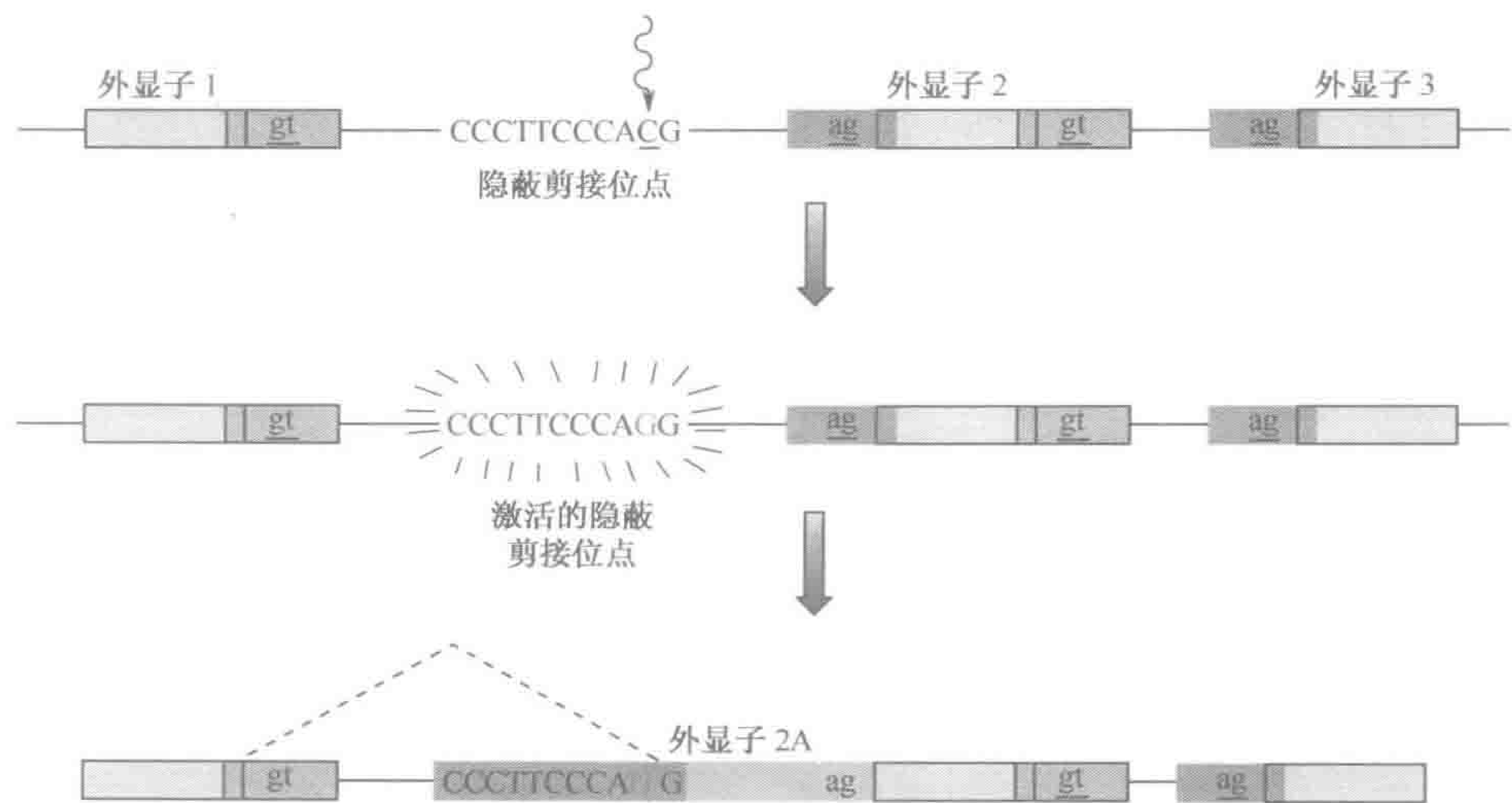


图 11.13 突变可通过激活隐蔽剪接位点导致不正常的 RNA 剪接

位于内含子中的一个隐蔽剪接受体序列的激活（与图 11.12 所示的外显子中隐蔽剪接供体的激活相对照）。突变可导致对于 RNA 剪接并不重要的序列的改变，以至于产生一个新的替代剪接位点。在所示的例子中，突变被设想改变了位于内含子 1 中的一个核苷酸。该核苷酸恰好位于一个与剪接受体一致序列非常相近、但并无保守性的 AG 双核苷酸的隐蔽剪接位点序列中（图 1.15）。突变消除了这种差别，将隐蔽剪接位点激活以至于它将与天然的剪接受体位点竞争。倘若它被剪接装置使用，一个新的外显子，即外显子 2A 便产生了，后者将含有额外的序列，可能会也可能不会导致移码。

产生干扰维持生命所必需的细胞功能的截短多肽所致的潜在致命后果。

▶ **截短多肽**（truncated polypeptide）。NMD 可确保截短多肽在活体中为一种稀有的产物，然而它将依赖于剪接。因此，5%左右缺乏内含子的人类基因（表 9.5）之一中

的无义突变将导致截短多肽。截短多肽的影响可能难以预测，并且将取决于其他因素如截短的程度、多肽产物的稳定性及其干扰正常等位基因表达的能力等。

- **外显子跳跃**——对一小部分无义突变来说，对于基因表达潜在的有害影响可被一个称为‘**无义相关的可变剪接**’（nonsense-associated altered splicing，NAS；Wang *et al*.,2002）的过程减轻。在这里，正常的剪接模式可通过诸如外显子跳跃等发生改变，以至于终止密码子被越过，产生一个没有突变的稳定 mRNA。NAS 的存在被一些人解读为肯定存在某种容许在 mRNA 转运到细胞质之前阅读密码子的核内翻译机制。然而，对于大多数，即使不是全部 NAS 事例的替代解释是该无义突变改变了**外显子剪接增强子**（exonic splice enhancer）[对于 NAS 如何被激活的其他观点见节 11.4.3 和 Maquat（2000）]。

11.5 重复序列的致病潜力

人类基因组中有很高比例的重复 DNA 序列（节 9.5 及 9.6），后者易于发生拷贝数变异和序列交换（表 11.6）。重复拷贝数的减少常可导致致病性缺失，但由序列复制所产生的扩延也能致病（Mazzarella and Schlessinger，1998）。特定染色体区域，尤其是亚端粒及着丝粒周边区域包含巨大的重复 DNA 束，而这类区域的不稳定性使疾病易于发生（Eichler，1998）。散在的重复序列也可以通过各种不同的机制产生致病性突变（表 11.6）。

表 11.6 重复 DNA 序列常导致疾病

重复 DNA 类型	突变类型	机制及例子
串联重复		
基因内非常短的重复	缺失	滑链错配(图 11.5)。图 11.14 中的例子
	移码插入	滑链错配
	三联重复的扩延	最初由滑链错配？之后通过未知机制进行大规模扩延
中等大小的基因内重复	基因内缺失	UEC/UESCE[a](图 11.7)
	部分或整体基因缺失	图 11.16 中的例子
	UEC/UESCE[a](图 11.7)	
包含完整基因的大型串联重复	基因序列的改变	图 11.16 和 11.17 中的例子
	基因转变(图 11.10)	
分散的重复		
短的正向重复	缺失	滑链错配或染色单体内重组？
分散重复元件(如 Alu 重复)	缺失/重复	UEC/UESCE[a](节 11.5.4)
倒位重复	倒位	染色单体内交换，如Ⅷ因子(图 11.20)
寡拷贝长重复	大范围缺失/重复	UEC/UESCE[a](表 11.7 和图 11.19)
活跃的转座元件	通过反转座子的基因内插入	反转座。例子见节 11.5.6

a UEC——不等交换；UESCE——不等性姐妹染色单体交换

11.5.1　短串联重复的滑链错配易于发生致病性缺失及移码性插入

由于通常将引起翻译移码，编码 DNA 内的插入和缺失较为罕见。然而，有时候，一个多肽的编码序列中偶尔也会存在一系列小数目核苷酸的串联重复序列。这类重复，如同微卫星基因座，相对容易因滑链错配而发生突变。结果，串联重复序列的拷贝数容易波动，从而引起一个或更多重复单位的缺失或插入。倘若突变发生在编码多肽的 DNA 中，所产生的缺失将常常对基因表达造成深刻的影响。移码性缺失通常将取消基因的表达并可能较为常见。

即使缺失未导致移码，一个或更多氨基酸的丢失仍可能致病（图 11.14）。小的移码性插入预期也将导致基因表达的丢失，并且这种插入经常属于其旁侧序列的串联重复。然而，非移码性插入通常应不会致病，除非这种插入发生在至关重要的区域中，使某种重要结构变得不稳定或者以某种方式妨碍了基因的功能。

11.5.2　短串联重复序列的不稳定扩延可导致各种疾病，但突变的机制并不十分清楚

基因内部或紧邻区域中的某些短串联重复可扩延至可观的长度并影响基因表达而导致疾病。有时一个导致疾病的中度扩延的重复可能十分稳定，并以不变的长度在若干世代中传播。然而，在其余的情况下，扩延的重复并不稳定。有关人类疾病可由非常不稳定的三核苷酸重复的大规模扩延的发现相当出人意料（对其他有机体的研究并未显示这一现象的先例），而人类例子的名单现已相当可观（表 16.6）。

尽管 64 种潜在的三核苷酸序列均具有可能，在考虑到循环排列 $(CAG)_n=(AGC)_n=(GCA)_n$ 以及自任意一条链阅读［一条链上的 5′(CAG)＝另一条的 5′(CTG)］时，在基因组 DNA 水平上仅有 10 种不同的三核苷酸重复（图 11.15）。除不稳定的三联重复扩延之外，位于导致进行性肌阵挛癫痫症的 cystatin B 基因上的大多数致病等位基因均涉及一个 12 核苷酸重复 $(C)_4G(C)_4GCG$ 的扩延（Lalioti *et al.*, 1997）。如表 16.6 所详述，包含不稳定扩延的短串联重复序列的基因根据扩延的大小和所在位置分为以下两大类：

- **导致多聚谷氨酸束** $(CAG)_n$ **的适度扩延**。密码子 CAG 编码谷氨酸。稳定的非致病范围为 10～30 个重复。不稳定的致病等位基因常具有 40～200 个重复。扩延的多聚谷氨酸束将导致蛋白质聚集在特定的细胞中并杀死它们；
- **极大的非编码重复扩延**。见于非编码序列（启动子、非翻译区或内含子序列）中的各种类型的重复（如 CGG、CGG、CTG、GAA）曾经历极大的扩延。这种扩延将抑制紧邻基因的表达，导致功能丢失。稳定的非致病等位基因有 5～50 个重复；不稳定的致病等位基因则具有数百或数千个拷贝（表 16.6）。**注意**：一些短的非编码性串联序列的极大扩延将仅仅影响染色体结构，形成**脆性位点**（fragile site），但不引起疾病（大概是由于附近没有重要的基因）。例子包括 FRAXF 和 FRA16A（均缘于 CCG 扩延）。

在各种情况下，小于特定阈值长度的重复在有丝分裂与减数分裂中是稳定的，但在大于阈值长度时则变得极端不稳定。这些不稳定重复实际上在从亲代到儿女的传递中将绝不会保持不变。扩延和缩短均会发生，但对于扩延有所偏倚。平均的长度变化经常将

取决于传递亲代的性别及重复的长度。

扩延机制的性质仍不十分清楚（Djian，1998；Sinden *et al*.，2002）。鉴于间断的重复似乎稳定而仅均质性重复不稳定的观察，滑链错配（图 11.5）被认为是扩延机制可能的组件。例如，在脊髓小脑共济失调症 1 型中，123/126 正常大小的 CAG 重复被一两个 CAT 三联碱基所间断，而 30/30 扩延的（CAG）$_n$ 等位基因则不包含间断（Chung *et al*.，1993）。但为何长度改变偏向于扩延尚不清楚。对于不稳定重复扩延的理解正在取得迅速的进展，推荐读者阅读最新的综述以获取更多信息。

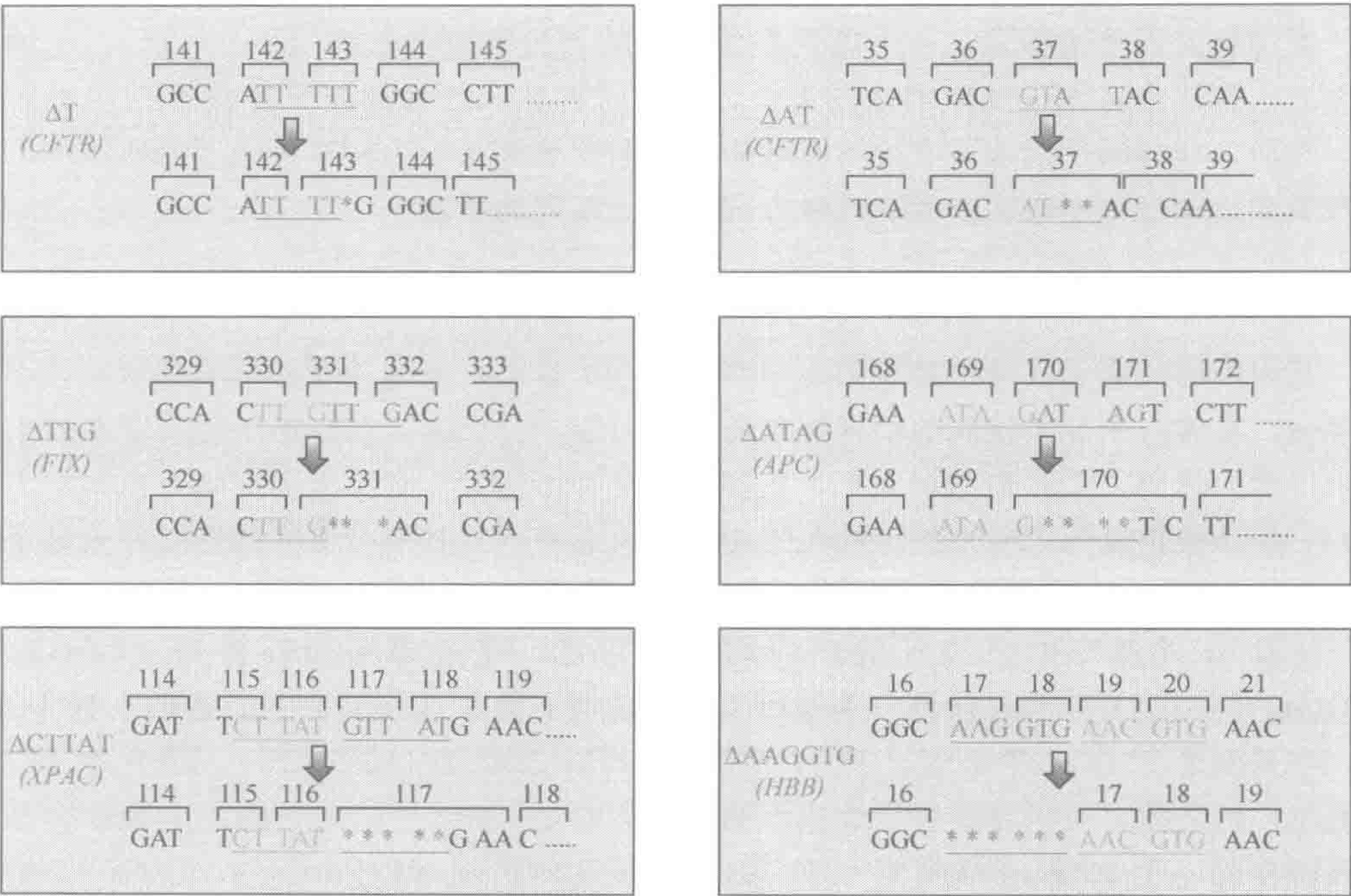


图 11.14　短串联重复是缺失/插入的热点

所示的 6 个缺失为发生在 1～6 bp 的串联重复单位处的致病性缺失的例子。3 bp 与 6 bp 的缺失不导致移码，其致病机理被认为是由于除去了 1 个或 2 个对于多肽功能至关重要的氨基酸。**注意**：在 6 bp 的缺失中，原始的串联重复并不完美。基因（以及相关疾病）为：*CFTR*，囊性纤维化跨膜调节因子；*F Ⅸ*，因子Ⅸ（乙型血友病）；*APC*，腺瘤样结肠息肉；*XPAC*，着色性干皮病 C 亚型；*HBB*，β珠蛋白（β地中海贫血）。尽管未在此显示，小型插入常为它们旁侧的串联重复序列。

AAC/GTT　AGG/CCT
AAG/CTT　ATC/GAT
AAT/ATT

ACC/GGT　CAG/CTG
ACG/CGT　CCG/CGG
ACT/AGT

图 11.15　10 种可能的三核苷酸重复

两条 DNA 链均见图示。所有其他的三核苷酸重复均属于这些重复中这个或那个的循环排列（见正文）。

11.5.3　串联重复且成簇的基因家族可能易于发生致病性不等交换及基因转变样事件

许多人类和哺乳动物基因簇中含有与功能基因密切相关的无功能性假基因。假基因与功能基因之间发生的基因座间序列交换可通过删除或改变功能基因的部分或全部序列而导致疾病。例如，功能基因与相关假基因间的不等交换（或不等性姐妹染色单体交换）将导致功能基因的缺失或含有源自假基因的片段的融合基因的形成。另外，假基因可作为基因交换事件中的供体序列，将有害突变引入功能基因中。

基因-假基因间交换致病的最经典例子为类固醇 21 羟化酶缺陷。其中 95%的致病性突变源于功能性 21 羟化酶基因 *CYP21B* 与一个关系甚密的假基因 *CYP21A* 间的序列交换。这两个基因被发现于约 30 kb 长的串联重复 DNA 片段上，后者亦含有其他的重复基因，特别是 C4 补体基因 *C4A* 和 *C4B*。大的致病性缺失都将造成含有 *CYP21B* 基因序列的约 30 kb DNA（对应于一个重复单位的长度）的去除（图 11.16）。相同长度的非致病性缺失也能找到，因为有时 *C4A* +*CYP21A* 重复单位将发生缺失（因而功能性 *CYP21B* 基因被保留下来），而 C4/21-OH 基因则表现为一种形式的大规模 VNTR 多态性基因座，不同的等位基因含有 1、2、3 或 4 个上述 30 kb 单位（Collier *et al.*，1989）。

事实上全部 75%的致病性点突变均拷贝自假基因中的有害突变，提示了一种基因转变机制（图 11.6 和 11.17）。对于一种新发生的此类突变的分析提示转变区域最大为 390 bp(Collier *et al.*,1993)。基因转变也见于重复的 C4 基因，二者均正常表达。在 *CYP21-C4* 基因簇内的转变中的一个可能的初期事件是染色单体不等配对，因此 *CYP21A-C4A* 单位将与 *CYP21B-C4B* 单位配对（图 11.17）。

11.5.4　分散的重复通常易于发生大的缺失与重复

短的直接重复

在几种情况下，缺失的末端的标志为非常短的直接重复。例如，线粒体 DNA 中众多的致病性缺失的断裂点均出现在完美或接近完美的短的直接重复序列上。其中，最常见的为一个 4977 bp 的缺失，发现于多名具有 Kearns-Sayre 综合征，一种以眼外肌麻痹、睑下垂、共济失调和白内障为特征的脑病的患者中。缺失将导致介于两个完美的 13 bp 重复之间序列的删除以及其中一个重复序列的丢失（图 11.18）。线粒体基因组缺乏重组，而 Shoffner 等（1989）曾推测这类缺失由一种复制滑脱机制引起，与发生于短串联重复处的情形相似（图 11.5）。线粒体基因组的部分复制亦是某些特定疾病的特征，尤其是 Kearns-Sayre 综合征。与普通缺失相似，重复序列的末端常以短的直接重复为标志，而复制与缺失的机制似乎密切相关（Poulton and Holt，1994）。

Alu 重复作为一个重组热点

一些大范围的缺失和插入可能由非等位性分散重复配对、随后是染色单体片段的断裂和重接而产生。例如，Alu 序列大约每 3 kb 就出现一次，而这类重复之间的错配被提出为缺失和插入的一种常见原因。一些大的基因在其内含子或非翻译序列中具有许多

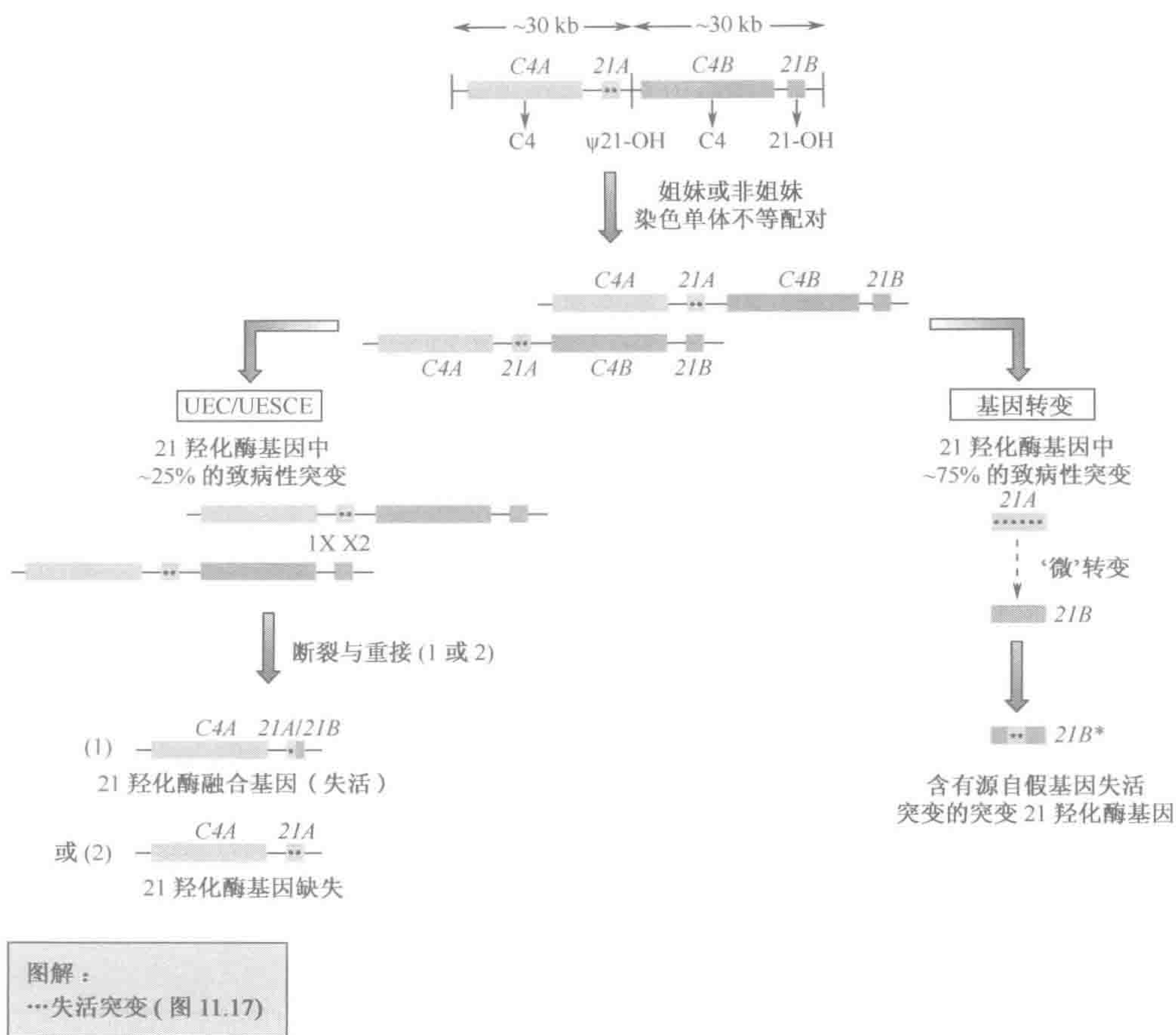


图 11.16　几乎所有的 21 羟化酶基因突变均缘于与非常相近的假基因的序列交换

重复的补体 C4 基因与类固醇 21 羟化酶基因位于串联的 30 kb 重复上，后者呈现约 97%的序列一致性。*C4A* 和 *C4B* 基因均表达产生补体 C4 产物；*CYP21B* 基因（*21B*）编码一种 21 羟化酶产物，而 *CYP21A*（*21A*）基因则是一个假基因。21 羟化酶基因座上约 25%的致病突变涉及由不等交换（UEC）或不等性姐妹染色单体交换（UESCE）所产生的一个 30 kb 缺失。其余的突变则为点突变，其中存在 *CYP21B* 基因的小规模基因转变——*CYP21A* 基因——一个包含有害突变的小片段被拷贝并插入到 *CYP21B* 基因中，替换了原始序列的一小段（一种可能的机制见图 11.10C)。基因转变事件，像 UEC 和 UESCE，很可能是由姐妹或非姐妹染色单体上的串联重复不等配对开始的。

内在的 Alu 序列，使它们易于发生频繁的内部缺失与重复。例如，在 45 kb 的低密度脂蛋白受体基因中每 1.5 kb 就存在一个 Alu 重复。该基因中极高频率的致病性缺失很可能涉及一种 Alu 重复，通常同时在两个末端，而偶然的致病性基因内重复亦涉及 Alu 重复（Hobbs *et al.*,1990)。这一观察表明了 Alu 序列在促进重组和重组样事件中的广泛作用。在成簇的多基因家族的进化中，最初的基因复制可能经常涉及 Alu 重复或其他散在重复元件之间的不等交换。然而，值得注意的是，一些富含 Alu 的基因看起来并非频发的 Alu 介导的重组的基因座。

突变位置	正常的 21 羟化酶基因序列 (*CYP21B*)	21 羟化酶假基因序列 (*CYP21A*)	突变的 21 羟化酶基因序列
内含子 2	CCCACCTCC	CCCAGCTCC	CCCAGCTCC
外显子 3 (密码子 110-112)	GGA GAC TAC TC Gly Asp Tyr Ser	G(.........)TC	G(.........)TC Val
外显子 4 (密码子 172)	ATC ATC TGT Ile Ile Cys	ATC AAC TGT	ATC AAC TGT Ile Asn Cys
外显子 6 (密码子 235-238)	ATC GTG GAG ATG Ile Val Glu Met	AAC GAG GAG AAG	AAC GAG GAG AAG Asn Glu Glu Lys
外显子 7 (密码子 281)	CAC GTG CAC His Val His	CAC TTG CAC	CAC TTG CAC His Leu His
外显子 8 (密码子 318)	CAC CAG GAG His Gln Glu	CAG TAG GAG	CTG TAG GAG Leu STOP
外显子 8 (密码子 356)	CTG CGG CCC Leu Arg Pro	CTG TGG CCC	CTG TGG CCC Leu Trp Pro

图 11.17　类固醇 21 羟化酶基因内的致病性点突变起源于自 21 羟化酶假基因拷贝序列

该拷贝被认为涉及一种基因转变样机制（图 11.16 和图 11.10C）。

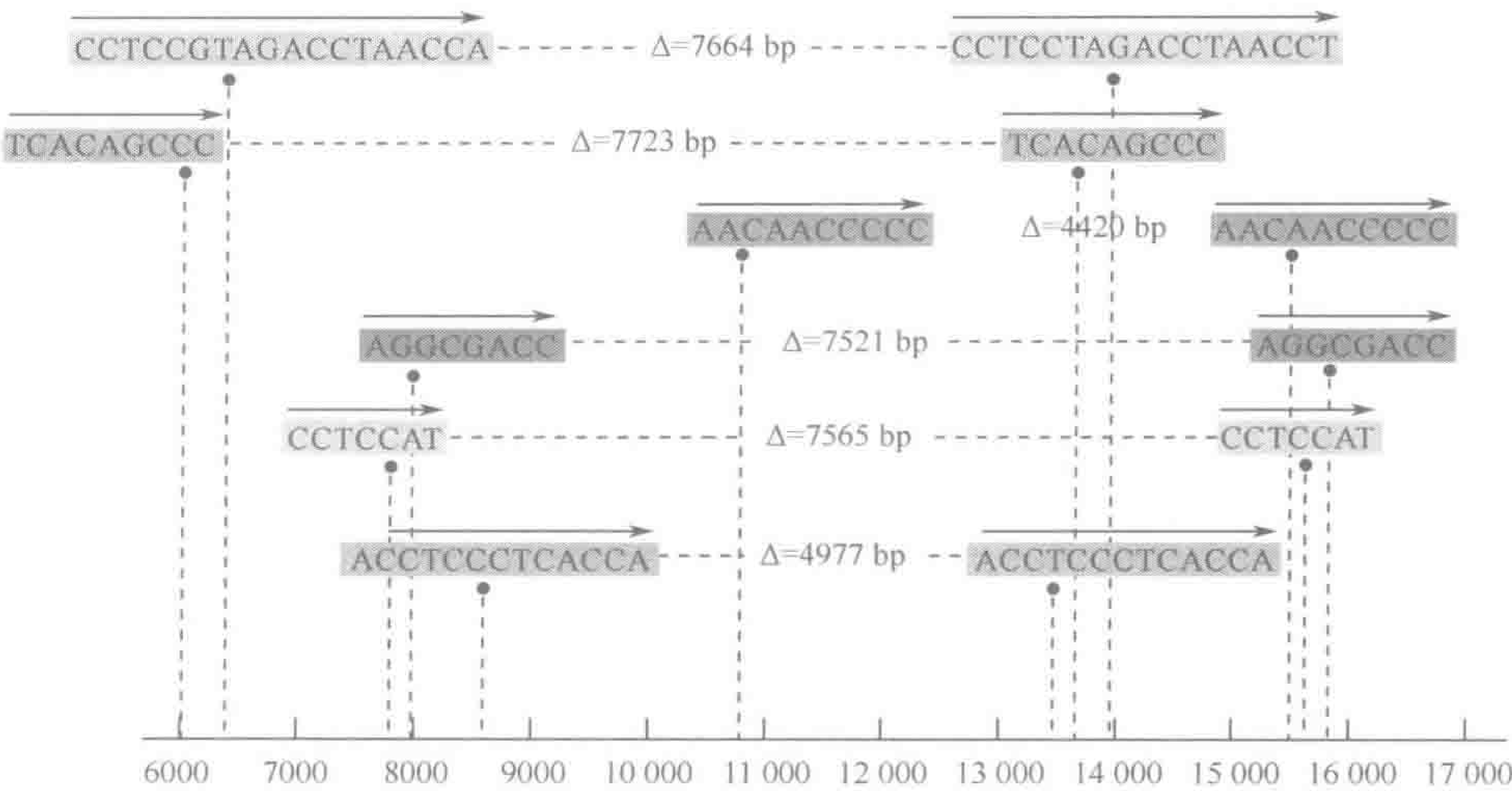


图 11.18　短的直接重复标志着线粒体基因组中许多致病性缺失的末端

注：由于线粒体缺乏重组，解释缺失的一种可能机制是滑链错配（见正文）。

低拷贝数的长重复

人类基因组的常染色质部分的测序揭示了非常多的散在低拷贝数序列的例子，后者显示出极高的序列同源性（序列一致率常＞95%），且常延伸超过数万至几十万对碱基。这种类型的重复序列通常曾经历进化中新近发生的、灵长类特异性的扩增，称为**节段性重复**（segmental duplication，节 12.2.5）。跨越如此长区域的极高序列同源性使重复（也称**重复子**，duplicon）之间的不等交换易于发生。当出现于不同染色体上时，这些

密切相关的重复可能使易位易于发生。当出现于同一条染色体上时，它们常使不等（=非等位基因性）同源重组易于发生，导致大范围的缺失和重复。

跨越超过百万碱基的间隔的重复子介导的缺失，以及较轻程度上的重复常常为致病性的（Stankiewicz and Lupski，2002；表 11.7 和图 11.19)，造成基因功能的丢失或间隔内剂量敏感性基因不适当的高基因剂量。这些重排通常不能由标准的细胞遗传学分析来分辨，并常被划分为（从染色体的角度看）微缺失及微重复。由于节段性重复在人类基因组中的广泛存在，重复子在发病机理中的作用可能相当大（Stank-iewicz and Lupski，2002)。

表 11.7　低拷贝数重复序列易于发生大范围病理性缺失和重复

性状/综合征	基　因	染色体定位	重排类型	大小(kb)	重复大小(kb)
男性不育(AZFa)	DBY,USP9Y	Yq11.2	Del	800	10
男性不育(AZFc)	RBMY,DAZ?	Yq11.2	Del	3500	229
Williams Beuren 综合征	ELN,GTF21?	7q11.23	Del	1600	320
Prader-Willi 综合征	?	15q12pat	Del	3500	500
Angelman 综合征	UBE3A	15q12mat	Del	3500	500
Smith Magenis 综合征	?	17q11.2	Del	3700	200
DiGeorge/VCF 综合征	TBX1,?	22q11.2	Del	3000/1500	225～400
Peripheral neuropathy(CMT1A)	PMP22	17p12	Dup	1400	24
Peripheral neuropathy(HNPP)	PMP22	17p12	Del	1400	24

经 Elsevier 授权，摘自 Stankiewicz and Lupski(2002). *Curr. Opin. Genet. Dev.* 12，312～319。

11.5.5　反向重复序列间染色质内重组可以产生致病性倒位

偶尔，具有高度序列一致性的成簇的倒位重复可能位于基因内部或附近。倒位重复之间的高度序列相似性使重复通过一种涉及一条染色单体向其自身折回的机制而配对易于发生。随后发生于错配重复处的染色质断裂与重接将产生一个倒位，与用于产生部分免疫球蛋白 κ 轻链的自然机制极为相似。

致病性倒位的经典例子是一种导致了 40%以上严重性甲型血友病的突变。Ⅷ因子基因 F8 的第 22 内含子含有一个 CpG 岛。两个内部的基因由此转录：F8A 与宿主基因 F8 的方向相反，而 F8B 则与 F8 同向（见图 11.20)。F8A 属于一个基因家族，具有另外两个密切相关的成员位于 F8 基因上游数十万对碱基处，并以与 F8A 相反的方向转录。因此，F8A 基因与其他两个成员之间的区域易于发生倒位——F8A 基因能与同一染色单体上的其他两个成员中的任何一个配对，随后于配对重复的区域内发生的染色单体断裂与重接将破坏Ⅷ因子基因（Lakich *et al*.,1993，图 11.20)。

几例由节段性复制产生的复制子（节 12.2.5）以及其他低至中等拷贝数的重复亦使倒位易于发生，但倒位并不直接致病，因为断裂点并未导致异常的基因表达（与Ⅷ因子基因中的例子不同)。相反，大范围的**倒位多态性**（inversion polymorphism）可能产生，如同位于 Williams-Beuren 基因座上的复制子（Osborne *et al*.,2001）以及高度同

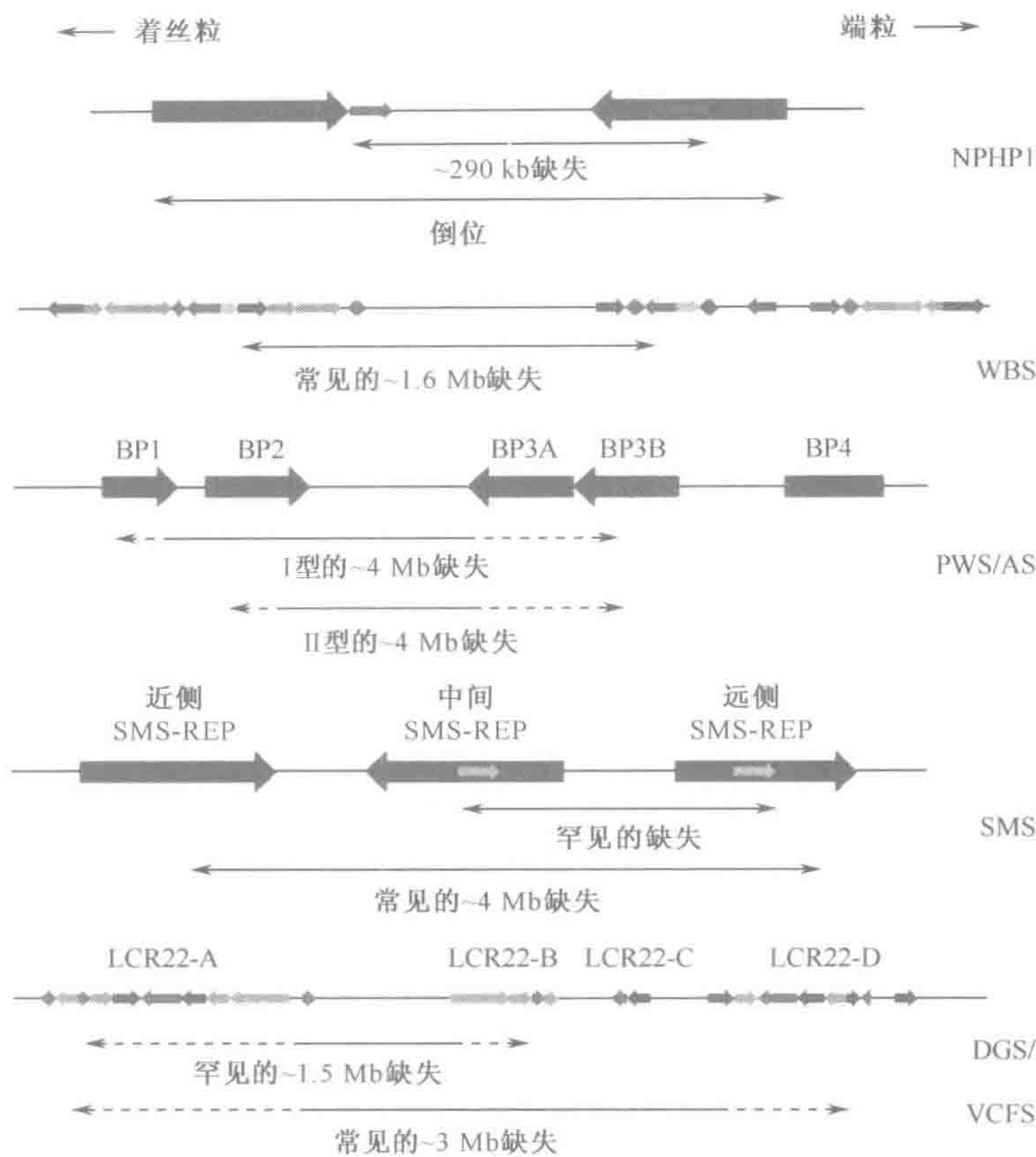


图 11.19 与人类疾病相关的选择性低拷贝重复序列（LCR）的复杂结构

注：LCR 的复杂结构由直接重复（同向箭头）和反向重复（反向箭头）组成。疾病名称缩写为：NPHP1：家族性青少年型肾结核Ⅰ型（familial juvenile nephronophthisis Ⅰ）；WBS：Williams-Beuren 综合征；PWS/AS：Prader-Willi 综合征/Angelman 综合征；SMS：Smith-Magenis 综合征；DGS/VFCS：DiGeorge 综合征/Velocardiofacial 综合征。经 Elsevier 授权，复制于 Stankiewicz 和 Lupski(2002). Trends Genet. 18，74～82。

源的嗅觉受体重复（Giglio *et al.*,2001）。

11.5.6 DNA 序列转座并非罕见，并能导致疾病

一部分中等及高度重复的散在元件能够通过一种 RNA 中介而发生转座（节 9.5）。由 DNA 转座所致的基因表达缺陷较为罕见，且仅代表分子病理学的一小部分。然而，若干由反转座所致的插入性失活造成的遗传缺陷的例子已见于记载。例如，在一项研究中，甲型血友病被发现于 140 名非近亲患者中的两人中，其原因为一个 LINE-1(Kpn) 重复被新插入到Ⅷ因子基因的一个外显子中。其他的例子据知为某个活跃转座的 Alu 元件所导致的插入性失活。此外，一些其他的例子记载为由不明 DNA 序列的基因内插入所致的发病机理。参考资料见 Kazazian(1998)。

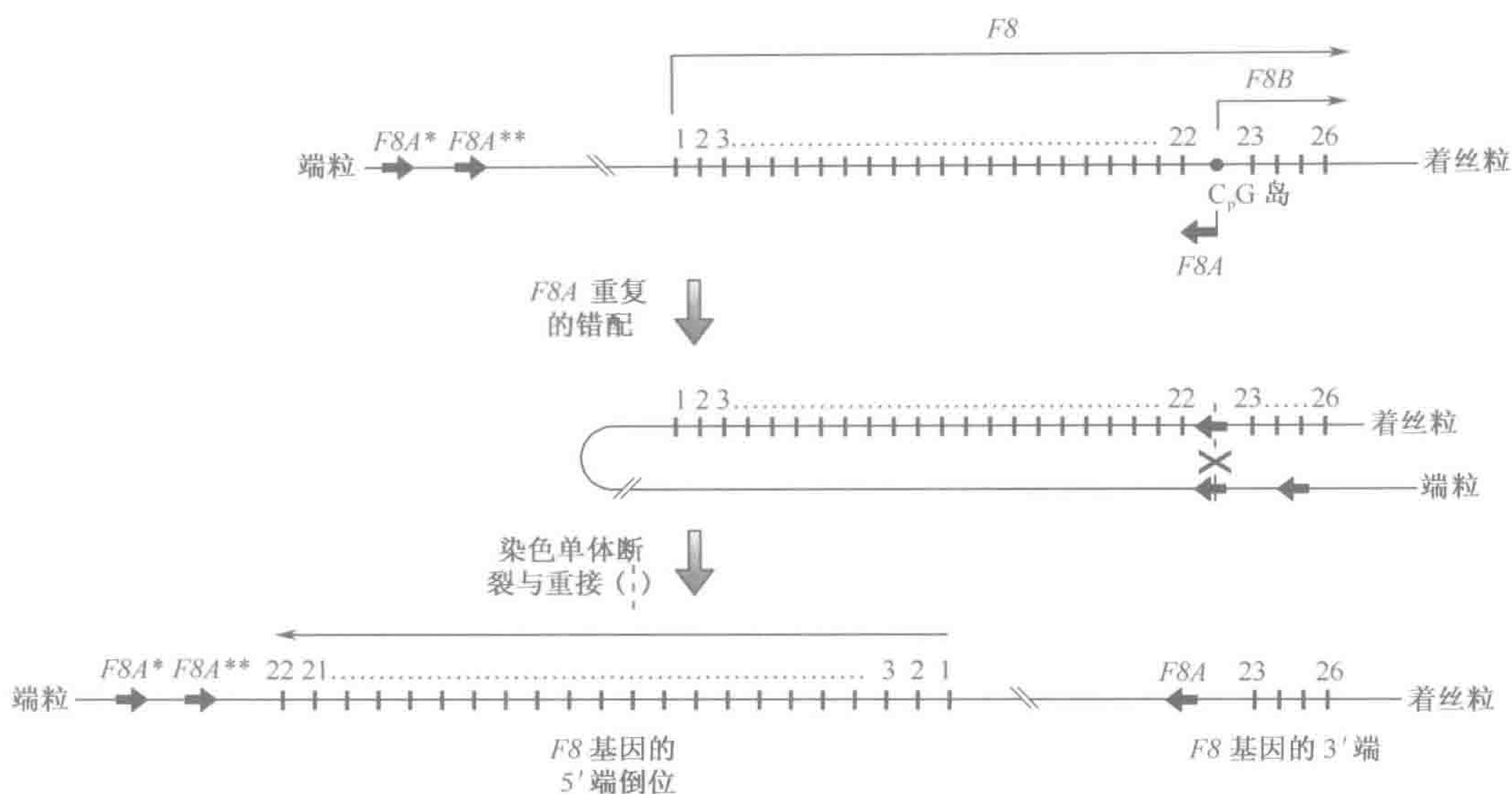


图 11.20 染色单体内倒置重复间的重组破坏Ⅷ因子基因的倒位

Ⅷ因子基因（*F8*）的第 22 内含子含有一个 CpG 岛，两个内部基因由此转录：同向的 *F8B*（一个新的外显子剪接到 *F8* 的第 23～26 外显子上）和转录自反向链的 *F8A* 基因。*F8A* * 和 *F8A* ** 为与 *F8A* 非常相近的序列，但位于上游约 500 kb 处且转录自反向链。*F8A* 基因家族三个成员之间高度的序列一致性意味着 *F8A* 基因与位于同一染色单体上的另两个成员之一的配对可通过染色单体折回而发生。随后的染色单体断裂和重接将导致 *F8A* 基因与配对的家族成员之间的区域发生倒位，造成Ⅷ因子基因的破坏（LaKich *et al.*,1993）。

11.6 DNA 修复

细胞内的 DNA 将受到各种各样的损伤，一部分可归因于细胞外的物质，但大部分是由于内源性机制，包括自发性化学水解，与活性氧基团在细胞内相互作用以及复制与重组错误等。

引起 DNA 损伤的细胞外因素

- **电离辐射**——γ 射线与 X 射线能引起 DNA 单链或双链断裂。
- **紫外线**——尤其是可被 DNA 强烈吸收的 UV-C 射线（～260 nm），但也包括能穿透臭氧层的波长更长的 UV-B 射线。紫外线能引起 DNA 链上相邻的胸腺嘧啶之间发生交联而形成稳定的化学二聚体（图 11.21C）。
- **环境化学物质**——包括烃类（如发现于烟雾中的烃类中的一部分），一些植物和微生物产物（如霉变的花生中产生的黄曲霉毒素）和用于癌症化疗的化合物等。**烷化剂**（alkylating agent）可将一个烷基（一个脂肪族的烃如甲基化基因）转移给碱基，并引起不同 DNA 链上或者一条链内碱基间的交联。

引起 DNA 损伤的内源机制

- **脱嘌呤**（图 11.21A）。通过碱基-糖链的自发裂解，每个有核人体细胞每天将丢失近

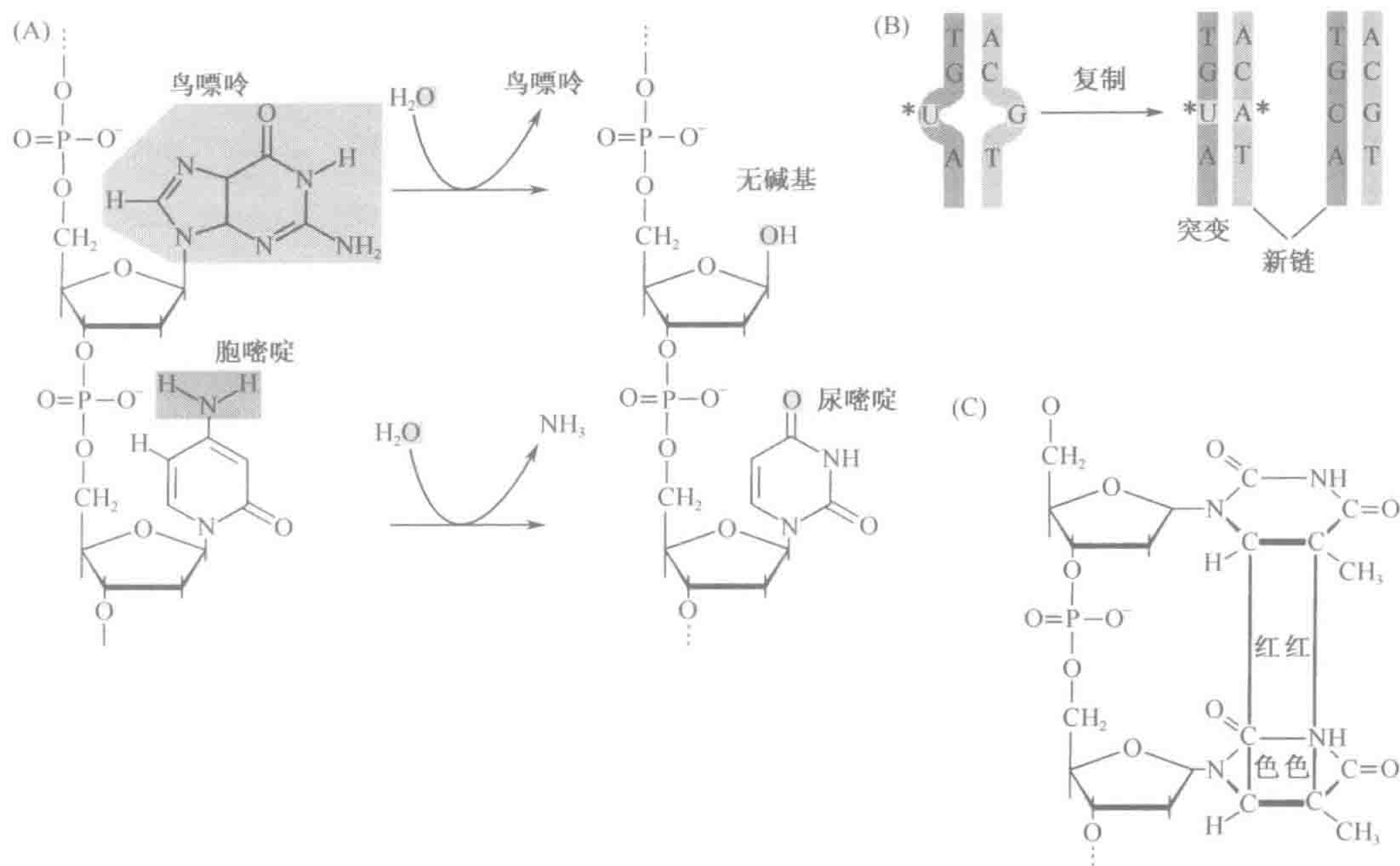


图 11.21　核苷酸的化学修饰导致的 DNA 损伤

(A) **脱嘌呤（上图）和脱氨基（下图）的例子**。(B) **化学修饰如何引起突变**。在这里胞嘧啶脱氨基产生尿嘧啶，倘若未被修复，DNA 复制将在互补链上插入一个腺嘌呤，因此净效应为 C→T 转换。通过碱基切除修复途径可能修复（图 11.21A）。脱嘌呤的净效应为缺失一个碱基被去除的核苷酸。(C) **胸腺嘧啶二聚体**——共价键（红色）连接相邻嘧啶碱基的碳原子。核苷酸切除修复途径可修复胸腺嘧啶二聚体（图 11.21B）。

5000 个腺嘌呤或鸟嘌呤。

- **脱氨基**（图 11.21A）。每个有核人体细胞每天都有约 100 个胞嘧啶自发脱氨基产生尿嘧啶（后者将优先与腺嘌呤碱基配对，导致 DNA 复制机制在遇到模版链上的 U 时将插入一个 A）。较少见的情况是腺嘌呤自发脱氨基产生次黄嘌呤。
- **活性氧类**。活性氧类包括过氧化阴离子 O_2^-，它既是一个离子又是一个**自由基**（free radical，一簇原子中有一个在最外面的电子层含有一个未配对的电子；这是一种极端不稳定的构象，自由基可迅速与其他分子或自由基发生反应而形成外层具有 4 对电子的稳定构象）。活性氧类通过作用于一些细胞分子上的电离辐射效应而产生，但也是细胞呼吸不可避免的的副产物（一些通过呼吸链向“下”传递的电子将偏离主要路径而直接去还原氧离子成为过氧阴离子）。在细胞内活性氧类将攻击嘌呤和嘧啶环。
- DNA **复制中的错误**。不正确的校读将导致碱基错配，例如尿嘧啶常常被不正确地插入到 DNA 中以取代胸腺嘧啶。
- **复制或重组的错误**将导致链断裂被遗留在 DNA 中。

倘若细胞要生存，所有这些细胞病变必须被修复。DNA 修复甚少为简单地逆转导

致损伤的改变（**直接修复**，direct repair）。几乎总是含有受损核苷酸的一段 DNA 被切除而该缝隙被重新合成（**切除修复**，excision repair）所填补。近 130 种参与 DNA 修复的人类基因（Wood *et al*.,2001 及http://www.cgal.icnet.uk/DNA Repair Genes.html #DR 上的附录）以及具有修复系统缺陷的人所具有的严重疾病（下面）突出了有效 DNA 修复系统的重要性。

11.6.1　DNA 修复常包括切除和重新合成损伤周围的全部 DNA 区域

为了应付所有这些形式的损伤，人体细胞能进行至少 5 种类型的 DNA 修复（综述参见 1995 年 10 月的 *Trends in Biochemical Sciences*；以及 Lindahl and Wood，1999）。

直接修复——逆转 DNA 损伤

三个基因被与这种不常用的机制相联系。其中，了解最为透彻者编码 O^6 甲基鸟嘌呤 DNA 甲基转移酶，后者能够除去不恰当甲基化的鸟嘌呤上的甲基。（**注意**：在细菌中，胸腺嘧啶二聚体能通过一种依赖于可见光以及一种酶的光催化反应除去。然而，尽管哺乳动物具有光分解酶相关的酶，它们却将其用于非常不同的目的，即控制它们的生物钟；Van der Horst *et al*.,1999)。

碱基切除修复（Base excision repair，BER）——用糖苷酶除去异常碱基（图 11.22A）

BER 能纠正许多最常见类型的 DNA 损伤（达到每天我们体内的每个有核细胞产生 20 000 个碱基改变的水平）。我们至少拥有 8 个编码不同 **DNA 糖苷酶**（DNA glycosylase）的基因，每个负责识别和除去特定种类的碱基损伤（表 11.8）。碱基切除后，一个内切核酸酶，**AP 内切核酸酶**（AP endonuclease）和一个磷酸二酯酶切开丢失碱基处的糖-磷酸骨架，并除去糖-磷酸残基。空缺为 DNA 聚合酶重新合成所填补，留下的缺口由 DNA 连接酶Ⅲ封闭。同样的过程也被用于修复自发性去嘌呤。

表 11.8　人类 DNA 糖基化酶（Wood *et al*.,2001)

基因	酶	公布的主要碱基改变
UNG	尿嘧啶 N 糖基化酶	U
SMUG1	单链选择性单功能尿嘧啶 DNA 糖基化酶	U
MBD4	甲基化 CpG 结合域蛋白 4	U 或 CpG 序列处 T 相对于 G
TDG	胸腺嘧啶 DNA 糖基化酶	U,T 或乙醇化 C 相对于 G
OGG1	8-氧化鸟嘌呤-DNA-糖基化酶 1	8 氧化 G 相对于 C
MYH	突变 Y 同源体	A 相对于 8 氧化 G
NTHL1(*NTH1*)	n^{th}内切酶Ⅲ样 1	环状饱和或片段化的嘧啶
MPG	N 甲基化嘌呤 DNA 糖基化酶	3 甲基化 A,乙醇化 A,次黄嘌呤

核苷酸切除修复（Nucleotide excision repair，NER）——除去胸腺嘧啶二聚体和大的化学加成物（图 11.22B）

NER 与 BER 的不同之处在于使用不同的酶，即使只有一个异常碱基需要修复，它

的核苷酸与许多其他邻近的核苷酸被一并除去；也就是说，NER 将除去周围的一大块损伤。图 11.22 示意了这个过程。核苷酸切除修复的缺陷将导致一种常染色体隐性的**着色性干皮病**（xeroderma pigmentosum，XP；Lambert *et al.*,1998）。通过细胞融合研究已经鉴定了 XPA-XPG 7 个亚群。XP 患者对紫外线极度敏感。暴露于阳光下的皮肤将出现数以千计的斑点，其中许多将发展成为皮肤癌。

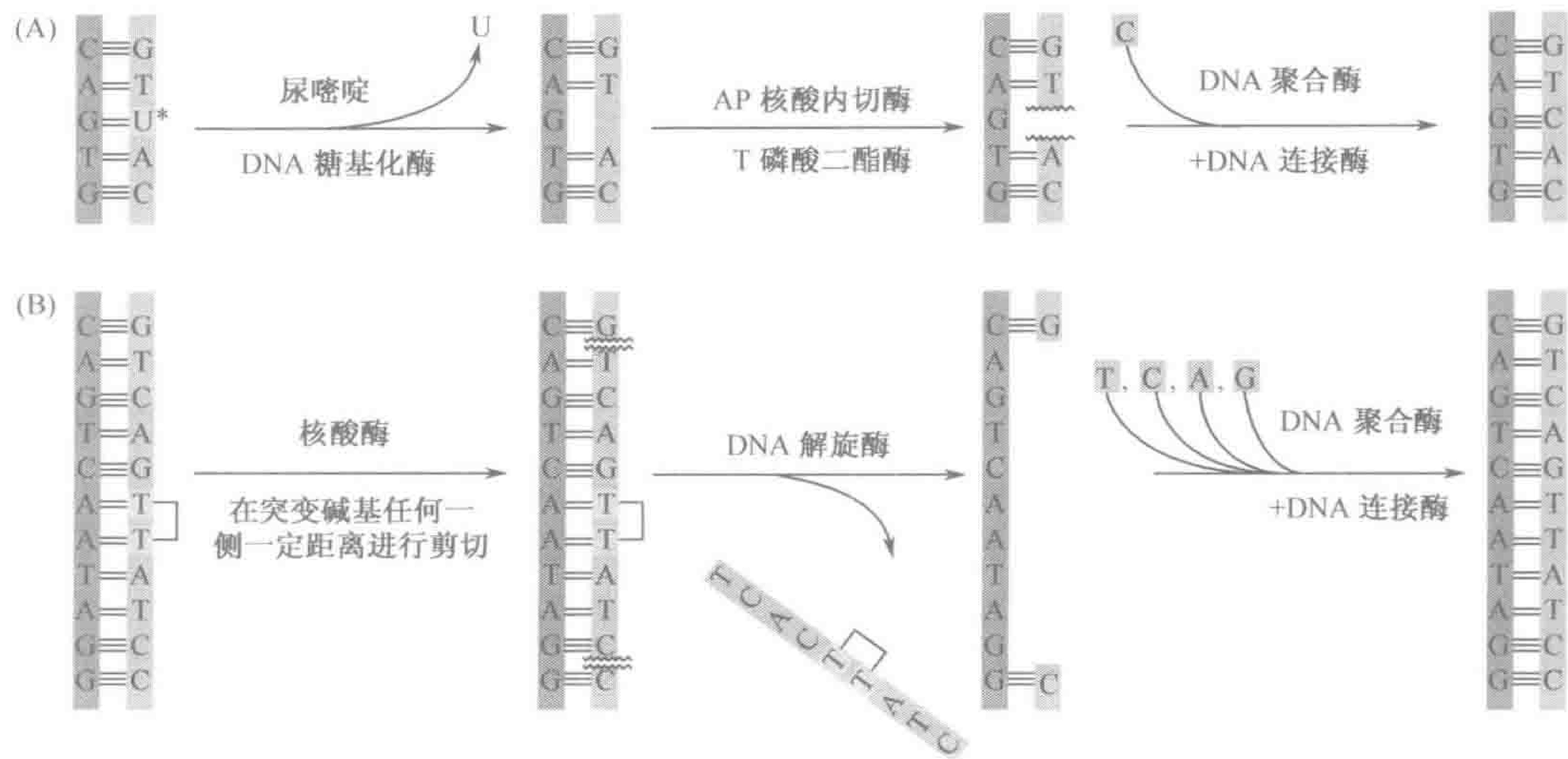


图 11.22　碱基和核苷酸切除修复途径

糖-磷酸骨架以垂直带阴影的框表示，氢键以⊜、⊜表示。

(A) **碱基切除修复**。在这里一个特异的 DNA 糖基化酶将一个胞嘧啶脱氨基后切除一个尿嘧啶（图 11.21A），残留的糖-磷酸碳水化合物部分随后被 AP 内切核酸酶和磷酸二酯酶相继除去。DNA 聚合酶将一个脱氧胞苷(dCMP) 插入，之后 DNA 连接酶将弥补这一缺口。(B) **核苷酸切除修复**。在这里嘧啶二聚体被识别为一个大的病灶；一个多聚酶复合体内的核酸酶在含有突变的链的任何一侧一定距离之外进行剪切，解旋酶将除去中间的片断产生一个大的缺口（实际上有 20 个以上的核苷酸，而不是这里所显示的小缺口）。该缺口由 DNA 聚合酶连续插入 dNMP 残基及随后的 DNA 连接酶封闭所修复。

纠正双链断裂需要复制后修复（内源重组）(Haber，1999)

通常的机制是一个基因转变样过程（**重组修复**，recombinational repair），来自同源染色体的一条单链侵入受损 DNA。除此之外，无论序列如何，断端都会重接在一起，这是一种可能引起突变的危险办法。对于重组修复的真核机制的了解不及切除修复系统清楚。涉及该通路的人类基因包括 *NBS*（突变见于 Nijmegen 断裂综合征；节 13.5.2），*BLM*（突变见于 Bloom 综合征；MIM 210900）以及乳腺癌易感基因 *BRCA2* 和 *BRCA1*（节 11.5.1）。

错配修复——纠正 DNA 复制错误所引起的错配碱基对

具有错配修复缺陷的细胞的突变率比正常细胞高 100～1000 倍，尤其在同源多聚体串内具有复制滑移的趋势（图 11.5）。在人类中该机制至少涉及 5 种蛋白质，其缺陷将

导致遗传性非息肉性结肠癌（节 17.5.3 和图 17.12）。

除直接修复外，所有这些系统均需要外切及内切核酸酶、解旋酶、聚合酶以及连接酶，通常以具有一些共同组件的多蛋白复合体形式发挥作用。整理出单独的途径在很大程度上得力于贯穿整个生命领域的修复机制的极强的保守性。不仅是反应机制而且是蛋白结构和基因序列从大肠杆菌到人类经常都是保守的。保守性的不利方面在于一套混淆的基因命名，有时是指人类疾病（*XPA* 等），有时是指酵母突变体（*RAD* 基因），有时又是指哺乳动物细胞互补系统（ERCC——切除修复交叉互补）：例如 *XPD*、*ERCC2* 和 *RAD3* 指的是人类、鼠类和酵母中的同一基因。一般说来，真核细胞对大肠杆菌的每套系统均具有多个对应的系统，因此，例如核苷酸切除修复在大肠杆菌内需要 6 种蛋白质，但在哺乳动物中则至少需要 30 种。

11.6.2 DNA 修复系统与转录和重组机构共享组件与过程

除彼此之间共享组件外，许多修复系统与 DNA 复制、转录和重组共享机构组件。在切除一个缺陷之后，DNA 聚合酶和连接酶是 DNA 复制和重新合成所必需的。这种重组机制与双链断裂修复有关。其与转录的联系尤其有趣（Lehmann，1995）。通用的转录因子 TFⅡH 为一个包含 XPB 和 XPD 蛋白的多蛋白复合体。TFⅡH 以两种形式存在。一种形式与普通的转录有关，另一种则与修复有关，且可能为转录活跃 DNA 修复所专有。该系统在两种罕见疾病中存在缺陷，Cockayne 综合征（CS；MIM 216400）以及毛发硫营养不良（TTD：MIM 601675）。临床上以及在细胞生物学方面，CS 和 TTD 均与 XP 存在重叠，并且在一些病例中相同的基因受累。然而，CS 和 TTD 患者具有可能反映了转录缺陷的发育缺陷，但他们没有 XP 患者所具有的癌症易感性。

11.6.3 对于损伤 DNA 物质的超敏性通常是受损细胞对 DNA 损伤反应的结果而非 DNA 修复缺陷

许多人类疾病包括对损伤 DNA 物质的超敏性或高水平的细胞 DNA 损伤并非由 DNA 修复系统自身的缺陷引起的，而是细胞对于 DNA 损伤反应的缺陷。

正常细胞通过将细胞周期的进展停滞在某个检查点上直至损伤被修复，或在损伤无法修复时触发凋亡的方式对 DNA 损伤产生反应。具有这种作用的部分机构包括 ATM 蛋白。ATM 蛋白的作用见节 17.5.1。简单地说，它将感知 DNA 损伤并将信号传递给“基因组卫士”P53 蛋白。缺乏功能性 ATM 的人将具有毛细管扩张失调症（MIM 208900；Lambert *et al.*,1998）。他们的细胞对辐射高度敏感，而且他们具有染色体不稳定性和发生恶性肿瘤的高风险，但 DNA 修复机构本身是完整的。Fanconi 贫血（MIM 227650）是另一组可导致对 DNA 损伤反应的缺陷、但并无 DNA 修复的特定缺陷的异质性疾病（至少 5 个亚群）。

（李春义　译）

进一步阅读

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111.

Cooper DN, Krawczak M (1993) *Human Gene Mutation.* BIOS Scientific Publishers, Oxford.

Graur D, Li W-H (2000) *Fundamentals of Molecular Evolution*, 2nd Edn. Sinauer Associates, Sunderland, MA.

Li W-H (1997) *Molecular Evolution.* Sinauer Associates Sunderland, MA.

Marnett LJ, Plastras JP (2001) Endogenous DNA damage and mutation. *Trends Genet.* **17**, 214–221.

Nickoloff JA, Hoekstra MF (eds) (1998) *DNA Damage and Repair. Vol. 2: DNA Repair in Higher Eukaryotes.* Humana Press, Totowa, New Jersey.

TIBS October 1995 issue on DNA repair (1995) *Trends Biochem. Sci.* **20**, 381–440.

电子资源与突变数据库

Codon Usage Database at http://www.kazusa.or.jp/codon/

Human Gene Mutation Database at http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html

Human DNA Repair Gene List at http://www.cgal.icnet.uk/DNA_Repair_Genes.html

HGV base (human genome variation database) at http://hgvbase.cgb.ki.se

Locus-specific mutation databases – see compilations at various centers e.g. within the Human Gene Mutation database at http://archive.uwcm.ac.uk/uwcm/mg/docs/oth_mut.html

Nomenclature for the description of sequence variations at http://www.dmd.nl/mutnomen.html

SNP databases – see compilations at various centers e.g. the UK HGMP Resource Centre at http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-mutation.html

参考文献

Ayala FJ (1999) Molecular clock mirages. *Bioessays* **21**, 71–75.

Blencowe BJ (2000) Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem. Sci.* **25**, 106–110.

Bodmer W, Cavalli-Sforza LL (1976) *Genetics, Evolution and Man.* Freeman, San Francisco.

Brown WM, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc. Natl Acad. Sci. USA* **76**, 1967–1971.

Cairns J (1975) Mutation selection and the natural history of cancer. *Nature* **255**, 197–200.

Chinnery PF, Thorburn DR, Samuels DC *et al.* (2000) The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends Genet.* **16**, 500–505.

Chung MY, Ranum LPW, Duvick IA, Servadio A, Zoghbi HY, Orr HT (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type 1. *Nature Genet.* **5**, 254–258.

Collier S, Sinnott PJ, Dyer PA, Price DA, Harris R, Strachan T (1989) Pulsed field gel electrophoresis identifies a high degree of variability in the number of tandem 21-hydroxylase and complement C4 gene repeats in 21-hydroxylase deficiency haplotypes. *EMBO J.* **8**, 1393–1402.

Collier PS, Tassabehji M, Sinnott PJ, Strachan T (1993) A de novo pathological point mutation at the 21-hydroxylase locus: implications for gene conversion in the human genome. *Nature Genet.* **3**, 260–265 and *Nature Genet.* **4**, 101.

Collins DW, Jukes TH (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386–396.

Cooper DN, Krawczak M, Antonorakis SE (2000) The nature and mechanisms of human gene mutation. In: *The Metabolic and Molecular Bases of Inherited Disease*, Vol. 1, 8th Edn (eds CR Scriver, AL Beaudet, WS Sly, D Valle). McGraw-Hill, New York.

Crow JF (2000) The origins, patterns and implications of human spontaneous mutation. *Nature Rev. Genet.* **1**, 40–47.

Djian P (1998) Evolution of simple repeats in DNA and their relation to human disease. *Cell* **94**, 155–160.

Dover GA (1995) Slippery DNA runs on and on and on. *Nature Genet.* **10**, 254–256.

Dubrova YE, Nesterov VN, Krouchinsky NG, Ostapenko VN, Neumann R, Neil DL, Jeffreys AJ (1996) Human minisatellite mutation rate after the Chernobyl accident. *Nature* **380**, 683–686.

Dubrova YE, Bersimbaev RI, Djansugurova LB *et al.* (2002) Nuclear weapons tests and human germline mutation rates. *Science* **295**, 1037.

Eichler EE (1998) Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8**, 758–762.

Ellegren H (2000) Microsatellite mutations in the germline. *Trends Genet.* **16**, 551–558.

Eyre-Walker A, Keightley PD (1999) High genomic deleterious mutation rates in hominids. *Nature* **397**, 344–347.

Fairbrother WG, Chasin LA (2000) Human genomic sequences that inhibit splicing. *Mol. Cell Biol.* **20**, 6816–6825.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB (2002) Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–1013.

Giglio S, Broman KW, Matsumoto N *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883.

Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* **185**, 862–864.

Grimm T, Meng G, Liechti-Gallati S, Bettecken T, Muller CR, Muller B (1994) On the origin of deletions and point mutations in Duchenne muscular dystrophy: most deletions arise in oogenesis and most point mutations result from events in spermatogenesis. *J. Med. Genet.* **31**, 183–186.

Haber JA (1999) Gatekeepers of recombination. *Nature* **398**, 665–667.

Hobbs HH, Russell DW, Brown MS, Golding JL (1990) The LDL receptor locus in familial hypercholesterolaemia: mutational analysis of a membrane protein. *Annu. Rev. Genet.* **24**, 133–170.

Hurst LD, Ellegren H (1998) Sex biases in the mutation rate. *Trends Genet.* **14**, 446–452.

International SNP Mapping Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Jeffreys A, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JAL (1994) Complex gene conversion events in germline mutation at human microsatellites. *Nature Genet.* **6**, 136–145.

Jenuth JP, Peterson AC, Fu K, Shoubridge EA (1996) Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA. *Nature Genet.* **14**, 146–150.

Kazazian HH Jr (1998) Mobile elements and disease. *Curr. Opin. Genet. Dev.* **8**, 343–350.

Kong A, Gudbjartsson DF, Sainz J *et al.* (2002) A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247.

Kumar S, Subramanian S (2002) Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808.

Lakich D, Kazazian Jr HH, Antonarakis SE, Gitschier J (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nature Genet.* **5**, 236–241.

Lalioti MD, Scott HS, Buresi C *et al.* (1997) Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* **386**, 847–851.

Lambert WC, Kuo H-R, Lambert MW (1998) Xeroderma pigmentosum and related disorders. In: *Principles of Molecular Medicine* (ed. Jameson JP). Humana Press, Totowa, New Jersey.

Lehmann AR (1995) Nucleotide excision repair and the link with transcription. *Trends Biochem. Sci.* **20**, 402–405.

Levinson G, Gutman GA (1987) Slipped strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.

Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326**, 93–96.

Li WH, Yi S, Makova K (2002) Male driven evolution. *Curr. Opin. Genet. Dev.* **12**, 650–656

Lindahl T, Wood RD (1999) Quality control by DNA repair. *Science* **286**, 1897–1905.

López Correa C, Brems H, Lázaro C, Marynen P, Legius E (2000) Unequal meiotic crossover: a frequent cause of *NF1* microdeletions. *Am. J. Hum. Genet.* **66**, 1969–1974.

Luo MJ, Reed R (1999) Splicing is required for rapid and efficient mRNA export in metazoans. *Proc. Natl Acad. Sci. USA* **96**, 14937–14942.

Lykke-Andersen J, Shu M-D, Steitz JA (2001) Communication of the position of exon–exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Science* **293**, 1836–1839 (see also the preceding paper in that issue).

Makova KD, Li WH (2002) Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416**, 624–626.

Maquat LE (2002) NASty effects on fibrillin pre-mRNA splicing: another case of ESE does it, but proposals for translation-dependent splice site choice live on. *Genes Dev.* **16**, 1743–1753.

Mazzarella R, Schlessinger D (1998) Pathological consequences of sequence duplications in the human genome. *Genome Res.* **8**, 1007–1021.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**, 297–304.

Osborne LR, Li M, Pober B *et al.* (2001) A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nature Genet.* **29**, 321–325.

Poulton J, Holt IJ (1994) Mitochondrial DNA: does more lead to less? *Nature Genet.* **8**, 313–315.

Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. *Trends Genet.* **16**, 296–302.

Reich DE, Schaffner SF, Daly MJ *et al.* (2002) Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genet.* **32**, 135–142.

Richard I, Beckmann JS (1995) How neutral are synonymous codon mutations? *Nature Genet.* **10**, 259.

Roy-Engel AM, Carroll ML, Vogel E *et al.* (2001) Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* **159**, 279–290.

Shoffner JM, Lott MT, Voljavec AS, Soueidan SA, Costigan DA, Wallace DC (1989) Spontaneous Kearns-Sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: a slip-replication model and metabolic therapy. *Proc. Natl Acad. Sci. USA* **86**, 7952–7956.

Sinden RR, Potaman VN, Oussatcheva E, Pearson CE, Lyubchenko YL, Shlyakhtenko LS (2002) Triplet DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J. Biosci.* **27**, 53–65.

Stankiewicz P, Lupski JR (2002) Molecular-evolutionary mechanisms for genomic disorders. *Curr. Opin. Genet. Dev.* **12**, 312–319.

Swallow DM, Gendler S, Griffiths B (1987) The hypervariable gene locus PUM, which codes for the tumour associated epithelial mucins, is located on chromosome 1, within the region 1q21-24. *Ann. Hum. Genet.* **51**, 289–294.

Takahara K, Schwarze U, Imamura Y *et al.* (2002) Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-alpha1(V) N-propeptides and Ehlers–Danlos syndrome type I. *Am. J. Hum. Genet.* **71**, 451–465.

Van der Horst GTJ, Muijtjens M, Kobayashi K *et al.* (1999) Mammalian CryI and Cry2 are essential for maintenance of circadian rhythms. *Nature* **398**, 627–630.

Vogel F, Motulsky AG (1996) *Human Genetics. Problems and Approaches*, 3rd Edn. Springer Verlag, Berlin.

Wang J, Chang YF, Hamilton JI, Wilkinson MF (2002) Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay. *Mol. Cell* **10**, 951–957.

Wood RD, Mitchell M, Sgouros J, Lindahl T (2001) Human DNA repair genes. *Science* **291**, 1284–1291.

Wu CI, Li WH (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl Acad. Sci. USA* **82**, 1741–1745.

Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW (2002) Allelic variation in human gene expression. *Science* **297**, 1143.

第 12 章　我们在生命之树中的位置

本 章 内 容

杜布赞斯基（Theodosius Dobzhansky）的观点‘除非从进化的角度来观察，否则任何生物学现象都将毫无意义’的支持者应已受到由各种基因组计划所涌出的大量数据的鼓励。从整个基因组范围进行序列比较现已成为**比较基因组学**（comparative genomics）这一新兴学科的主流研究内容，并且它们正开始为我们在生命之树中的位置提供有力的新线索。

当然，人类基因组（以及所有的基因组）的进化起源同生命本身一样古老，但本章并不准备对分子进化遗传学本身进行总结。因此，许多令人着迷的领域，诸如遗传密码子的早期进化（Knight and Landweber，2000）或有关在被 DNA 取代之前，RNA 曾是主要的信息分子的观点（Joyce，2002）等在这里并未被覆盖。

相反，本章准备将焦点集中在对于现存基因组的比较分析是如何来揭示人类 DNA 以及人类基因的进化起源的。许多数据均来源于对哺乳动物以及其他动物基因组的比较，尽管对于亲缘更远的基因组的比较偶尔也会被用来解释特定的**进化足迹**（footprint of evolution），诸如内含子与线粒体 DNA 的起源等。比较的数据将有助于了解我们区别于其他哺乳动物模型，特别是小鼠（了解人类早期发育以及人类疾病的一个重要模型）以及现存与我们亲缘最近的灵长类的独特之处。结尾的部分将考察人类种群的最近起源以及其间的亲缘关系。

12.1 基因结构与复制性基因的进化

真核生物的基因及蛋白质通常较那些源自简单有机体的大，且更为复杂。真核生物基因较大的尺度主要是缘于存在大的内含子，而内含子的平均长度通常可以反映基因组（以及生物学上）的复杂性。而由于不同的机制，真核生物基因的编码序列通常亦长于原核生物。基因内复制将使编码序列长度扩增并通常发生歧化。基因间重组则将促成**蛋白质结构域**（protein domain，离散的结构或功能模块）的不同组合。内含子在复杂基因组内的基因中近乎普遍的存在被认为反映了它们在进化过程中使编码序列得到扩增和修饰中的重要作用。

12.1.1 剪接体内含子很可能起源于第二组内含子并首次出现在早期的真核细胞中

在 1977 年分裂基因被发现之后，**剪接体内含子**（spliceosomal intron）（框 12.1）引起了激烈的争论。发现于复杂基因组中的内含子通常要长于其他物种，并且内含子序列并未得到很好的保留。尽管如此，内含子含有了涉及基因调控的重要功能序列，并且一些较短内含子的序列在进化中得到了相当程度的保留。强烈表达的基因通常具有非常短的内含子（可能是缘于对快速 mRNA 加工的选择）。无论对内含子的功能有何种推测（例如，像节 12.1.3 中那样容许重组以实现进化学上的新颖性），答案肯定不止一种：复杂有机体中的一小部分基因中没有内含子（表 9.5）。

框 12.1 内含子组

内含子为具有不同功能以及包括巨大长度差异（而不像外显子那样在长度上似乎要均匀得多，表 9.6）在内的显著结构差异的异质性实体。根据其在 RNA 剪接方面对外源性因素的依赖程度以及剪接反应的性质，它们可以被分为以下几个**内含子组**（intron group）：

- **剪接体内含子**（spliceosomal intron）为真核细胞中最常见的内含子。它们被转录为原始转录物中的 RNA，并在剪接体对 RNA 进行加工的过程中在 RNA 水平被切除。仅有几条短序列似乎对基因的功能具有重要性［位于或靠近剪接接口与位于分支点者（图 1.15）以及剪接增强子与沉默子（节 11.4.3）］。因此，剪接体内含子可能容许大型的插入且可能相当长（有时超过 1 Mb）。剪接体内含子很可能在进化中出现得相对较晚，并且可能来源于Ⅱ型内含子（节 12.1.1）。通过容许移动元件的插入，它们可协助外显子混编。
- **Ⅰ型和Ⅱ型内含子**（group Ⅰ and Ⅱ intron）具有突出的二级结构，并且为**自剪接内含子**（self-splicing intron）（它们能催化自身的切除而不依赖于剪接体）。它们被发现于细菌及真核生物中，但其分布却非常局限，主要见于 rRNA 与 tRNA 基因以及发现于某些类型的线粒体、叶绿体以及噬菌体中的少数蛋白质编码基因中。两类均可作为移动元件，可移动Ⅱ型内含子编码一种反转录酶样活性，与 LINE-1 元件出奇的相似。Ⅰ型和Ⅱ型内含子区别于保守的剪接信号及其剪接反应的性质（例如，仅Ⅰ型内含子需要一个游离的鸟嘌呤碱基；Bonen and Vogel，2001）。
- **藻类内含子**（archaeal intron）仅见于藻类的 tRNA 及 rRNA 基因中。与Ⅰ型和Ⅱ型内含子不同，它们没有保守的内部结构，不能自剪接。尽管其剪接机制依赖于蛋白质，但与剪接体内含子不同，它们的剪接反应并不需要具有顺式作用的 RNA 分子。

剪接体内含子的进化一直是存在争议的问题（Logsdon，1998；Lynch and Richardson，2002）。目前大体上被认同的是剪接体内含子并非起源于真核细胞出现之前（已测序的几百个原核生物基因组内的编码蛋白质基因中无一含有现存或原有内含子的痕迹）。相反，它们很可能首次出现于真核细胞进化的很早阶段：唯一未发现其存在的可能分化较早的真核生物群为诸如毛滴虫之类的副基体，然而后者却具有部分必要的剪接机构。

剪接体内含子可能起源于具有相似加工机制的自剪接Ⅱ型内含子（框 12.1；Lynch and Richardson，2002）。剪接体内含子并不能进行自剪接，而是需要五种独立的 snRNA 分子以及多种蛋白质，但Ⅱ型内含子却通常为黏性，是能进行自剪接的内含子。然而，某些功能性Ⅱ型内含子据知已断裂成为不同的片段，在一定程度上与散在的剪接体加工系统相似。某些Ⅱ型内含子亦编码其自身的反转录酶，并类似于移动元件。因此，剪接体可能起源于某种早期的细胞器中的一种Ⅱ型内含子（由细胞器 DNA 向核 DNA 的序列转移已经被证实，框 12.4）。随后的片段化为不同的成分可能花费了相当长的时间。

自其首次出现之后，剪接体内含子在进化中被周期性地整合入基因中（在某些时候也会被清除）。其中一些明显具有古老的起源。例如，在珠蛋白超基因家族中，两个主要内含子的位置得到了很好的保留，提示二者很可能在 8 亿多年之前即已整合入一个原始的珠蛋白基因中（图 12.1）。内含子位置极高的保守性亦可见于亲缘较远的种间同源基因中。例如，人类的 Huntington 舞蹈病基因具有跨越 170 kb 的 67 个外显子。其在河豚（与人类于 4 亿多年以前分离）中的相应基因仅长 23 kb，但却同样具有 67 个外显子以及内含子位置近乎完美的保留（Baxendale *et al.*,1995）。然而，一些其他的剪接体内含子似乎具有更近的进化起源。例如，在众多独立的基因家族（如肌动蛋白、肌珠蛋白、微管蛋白等）中，内含子的位置并未得到很好的保留。

12.1.2　复杂基因可通过基因内复制进化，且通常缘于外显子复制

同其他真核生物基因相似，人类基因常常呈现基因内 DNA 复制的证据，且可能相当丰富。例如，许多基因已知编码完全或大体上由大型重复序列所构成的多肽，其中部分基因中的重复序列之间具有极高的序列同源性（表 9.7）。通过对既往形成的结构域进行复制，可产生具有各种进化优势的更长的多肽。人类的泛素编码基因，*UBB* 与 *UBC*，即编码较长的由完整的泛素蛋白单元串联重复所构成的多聚体蛋白，后者通过裂解而产生多个泛素拷贝（*UBB* 有 3 个，*UBC* 有 9 个）。

除了奇特的泛素基因之外，基因内复制通常涉及某种形式的**外显子复制**（exon duplication），从而造成对某个蛋白质结构域的复制（图 12.2）。约 10%的人类、果蝇（*Drosophila melanogaster*）以及秀丽新小杆线虫（*Caenorhabditis elegans*）基因具有重复的外显子（Letunic *et al.*,2002），若干优势可以被预知：

- **结构的延伸**　对于具有重要结构性功能的蛋白质来说，结构域的重复可能具有特殊的优越性。一个明显的例子就是具有 41 个外显子的 *COL1A1* 基因，后者编码构成三螺旋的 α(I) 胶原蛋白的部分，每个外显子主要编码整倍的（1～3 份）由 Gly-X-Y（X 与 Y 为各种氨基酸）的六联重复所形成的 18 氨基酸基序。

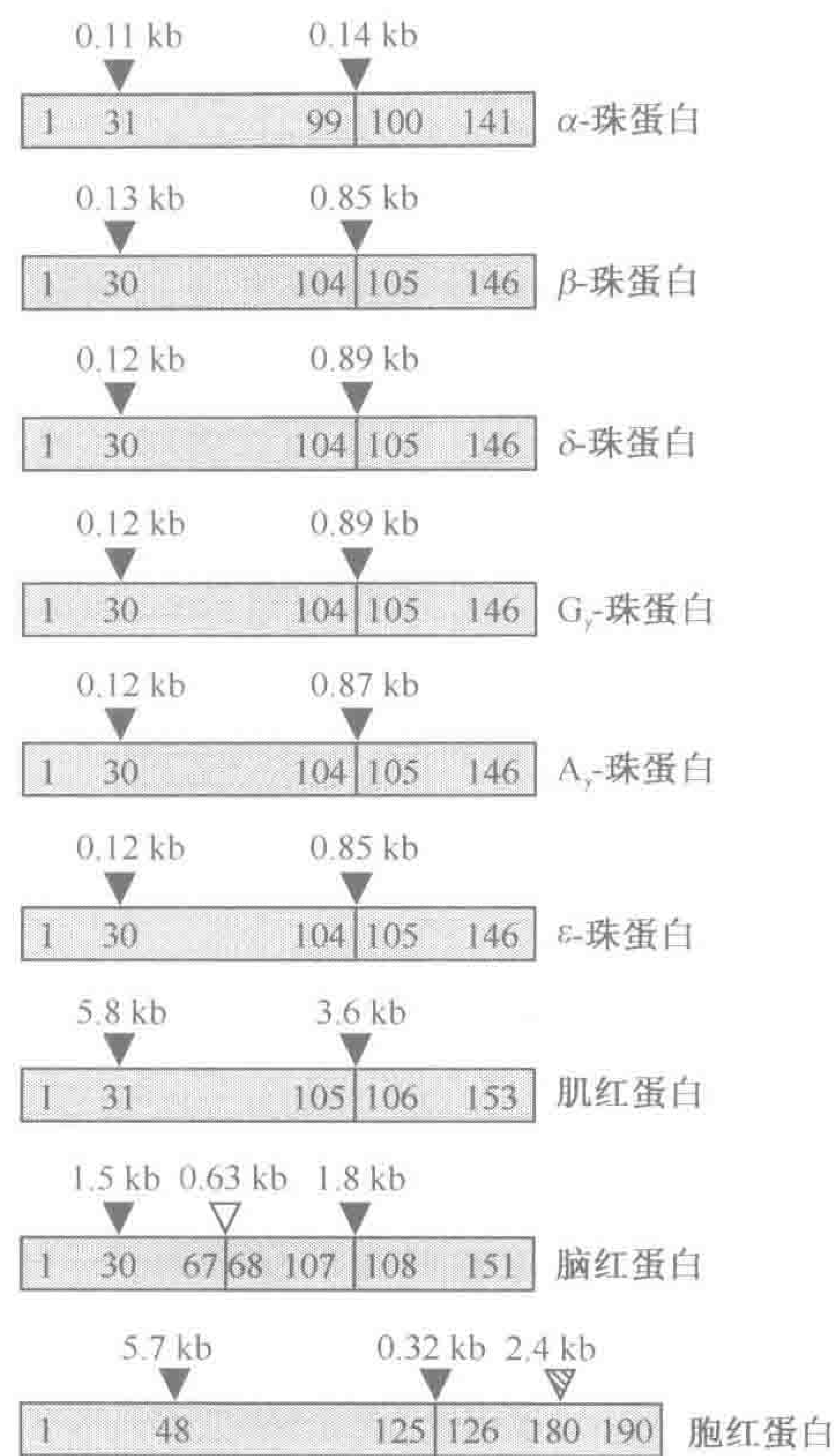


图 12.1　由珠蛋白超基因家族中内含子位置的保留所提示的古老的内含子插入

方框代表终产物多肽。框中的数字为氨基酸的位置。注意普遍存在的高度保守性提示在 8 亿多年以前（图 12.4），一条原始的珠蛋白基因具有两个内含子（▼）。最近发现的脑红蛋白序列（▽）以及胞红蛋白序列（▽）中额外的内含子似乎于相当长的时间之前被插入。注意胞红蛋白的扩张包括 N 端区域的 17 个氨基酸以及 C 端的 23 个氨基酸（Pesce *et al.*，2002）。

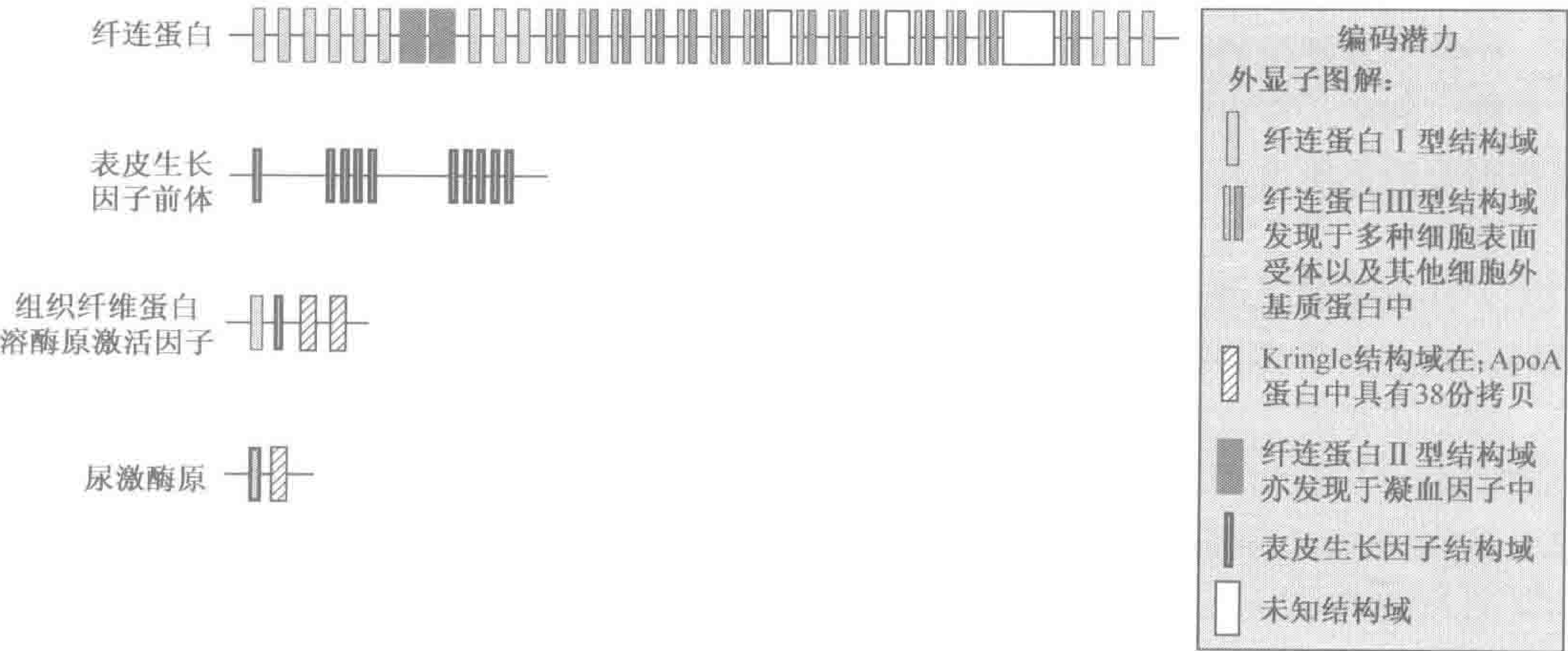


图 12.2　外显子复制与外显子混编

明显的外显子复制可见于人类纤连蛋白（*FN1*）与表皮生长因子（*EGF*）基因中。*FN1* 基因中具有一个编码Ⅰ型纤连蛋白的外显子的 12 份拷贝以及另一对外显子的 15 份拷贝。后两者共同编码Ⅲ型纤连蛋白结构域。*EGF* 基因具有一个编码 *EGF* 结构域的外显子的 9 份拷贝。外显子混编意味着编码这些结构域的外显子亦可见于诸如编码组织纤维蛋白溶酶原激活因子以及尿激酶原在内的其他许多基因中。外显子混编的可能机制见图 12.7。

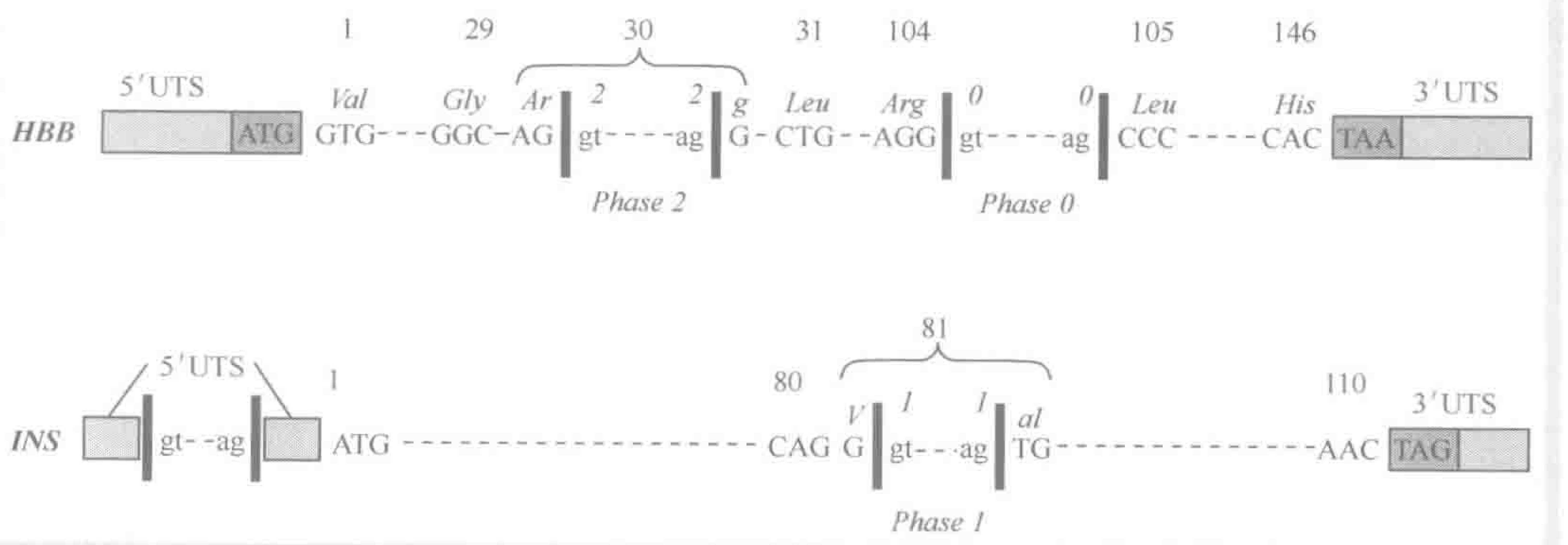
- **由结构域分化而形成的多样性**　在大多数情况下，随基因内复制事件而来的是各重复单元之间显著的核苷酸序列歧化。这种歧化将可能为获得相关但却不同的功能提供机会。有些时候重复序列之间的差异是如此之大，以至于在诸如免疫球蛋白结构域的例子中，在序列水平重复的结构已不明显（图 9.10）。
- **由选择性剪接而形成的多样性**（Letunic *et al*.,2002）。一类选择性剪接将通过从一组重复的外显子中选择一个外显子序列来放入剪接产物中，以产生不同的异构体。由于这些重复的外显子的序列可能具有微小的差异，这类剪接因而可能产生一系列相关的异构体。许多人类基因均进行这类剪接，但果蝇的 *Dscam* 基因却是最佳的例子（图 10.16）。

基因内的外显子复制可能由包括不等交换或不等姐妹染色体交换在内（Long，2001）的各种机制来解释。为了避免翻译读框的移码，复制仅局限于**对称性外显子**（symmetrical exon）或对称性外显子组中（框 12.2）。

框 12.2　对称性外显子与内含子相位

并非所有的外显子都含有编码 DNA。由于内含子有时位于非翻译序列如多肽编码基因的 5′UTR与 3′UTR 序列之中，因此存在**非编码外显子**（noncoding exon）。将编码序列分隔为不同编码外显子的内含子，通过外显子复制与外显子混编而容许歧化。根据其插入编码 DNA 的位点（见图），**内含子相位**（intron phase）可被区分为三类：0 相位（位于某个密码子的第三个碱基与下一个密码子的第一个碱基之间）；1 相位（位于某一密码子的前两个碱基之间）；2 相位（位于某一密码子的后两个碱基之间）。

编码外显子可根据其旁侧的内含子的相位分类。**对称性外显子**（symmetrical exon，核苷酸数目为 3 的整倍）将被同相位的内含子所包围，因而可能根据旁侧内含子的相位被区分为：0-0，1-1 或者 2-2 型。**非对称性外显子**（nonsymmetrical exon，核苷酸数目并非 3 的整倍）可以被分为 0-1，0-2，1-0 等类型。通过外显子复制和外显子混编来拷贝一个外显子通常仅限于对称性外显子，或一组不会因重复或拷贝入另一个基因而造成移码的相邻的非对称性外显子，例如相邻的 0-1 与 1-0 型外显子。



β 珠蛋白基因与胰岛素基因中内含子相位举例

基因名上方的数字代表密码子/氨基酸的位置。*HBB*——人类 β 珠蛋白基因。*INS*——人类胰岛素基因。**注意**：由于被不同相位的内含子所包围，即上游为一个 2 相位内含子，下游为一个 0 相位内含子，β 珠蛋白基因的第二外显子可被归类为非对称性的 2-0 型。

12.1.3　外显子混编将促成蛋白质结构域的新组合

在整个自然界中已知的保守蛋白质结构域仅有数千种，但是在后生物种中，许多蛋白质均含有见于另一种蛋白质中的结构域（Li *et al*.,2001）。纤连蛋白是一种大型的细胞外基质蛋白，含有由单个外显子或外显子对所编码的多个重复结构域，并且是外显子复制的经典例子。其中的重复结构域之一，目前称为纤连蛋白Ⅰ型结构域，被随后发现于组织纤维蛋白溶酶原激活因子中。与纤连蛋白相似，组织血浆蛋白酶原活化剂亦包含其他结构域。后者包括以表皮生长因子（EGF）前体为特征的类 EGF 结构域以及两个见于其他多肽如尿激酶前体以及纤维蛋白溶酶原等中的 kringle 结构域（图 12.2）。

由个别对称性外显子或对称性外显子组（框 12.2）来编码蛋白质结构域的情况并非罕见。这些发现提示了发生于基因之间的**外显子混编**（exon shuffling）的可能性：编码整个结构域的外显子或外显子组被复制并插入其他基因之中（Patthy，1999；Kaessmann *et al*.,2002）。外显子混编的机制最有可能涉及 LINE 元件协助的反转座（节 12.1.6）。

12.1.4　基因复制在多细胞有机体的进化中扮演了至关重要的角色

突变是进化的发动机。原核生物较短的倍增时间及较大的群落使带来生存优势的突变（因应环境的变化）能够迅速固定下来。然而，在多细胞动物中，重要基因的改变常常都是有害的，并且通常存在强大的保守性选择压力来保持一个基因的序列。因此，在复杂有机体中，突变的传播将需要通过**基因复制**（gene duplication）。

对每例基因复制来说，两个基因拷贝之一属于冗余，因此能够快速地歧化（因为缺乏保留其功能的选择压力）。即使不存在基因剂量效应的问题，额外的基因拷贝常常因发生有害的突变而蜕变成一个无功能的**假基因**（pseudogene），或者在 DNA 更替过程（当某段 DNA 序列并非至关重要时，可能需要通过漫长的进化时光来发生）中被丢失。然而，在某些情况下，分化的基因拷贝通过突变而产生可能造成选择优势的新性状的功能性产物而被保留下来（图 12.3）。编码序列的改变可能需要一些时间才能被固定下来，而调节序列的变化则可能迅速提供新的表达特征。所形成的基因产物可能表达于不同的组织中或者发育过程的不同阶段并随后适应于新的环境（见珠蛋白基因的例子，节 12.1.5）。

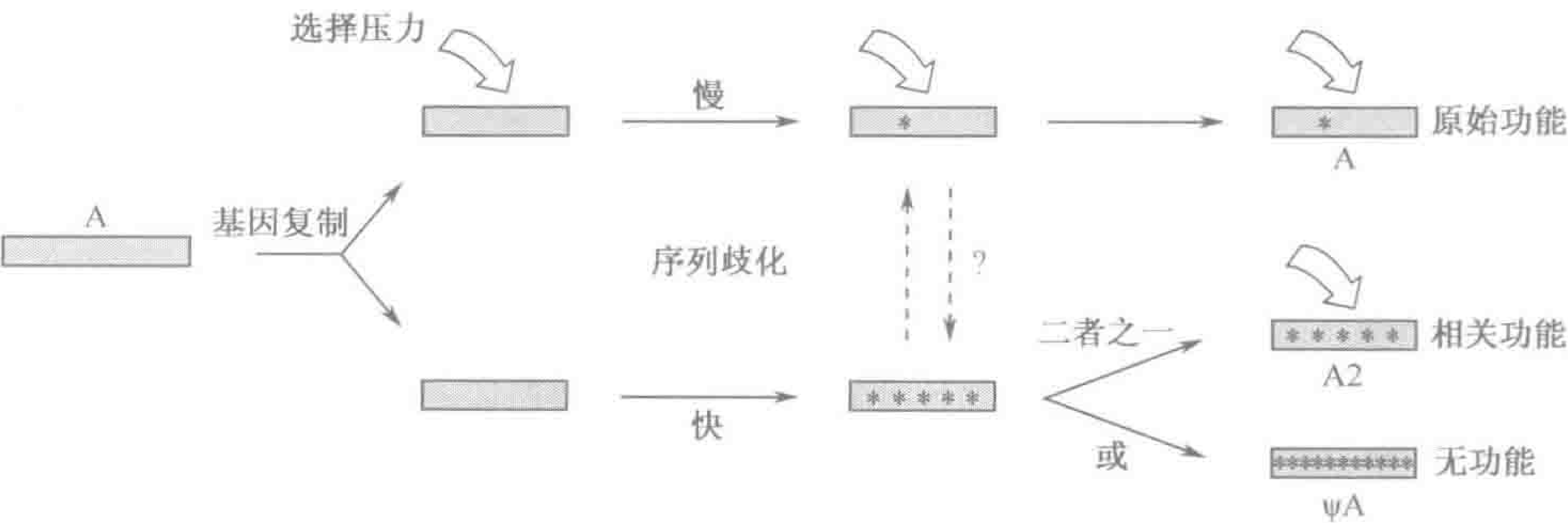


图 12.3　基因复制可产生不同的功能性基因变异但通常将导致假基因的形成

A 基因的复制将造成两个相同的拷贝。选择压力仅需要被施加于一个基因拷贝（上方）上，以维持原有功能性基因产物的存在。另一个拷贝（下方）将继续表达，但在并无选择压力以保持其序列的情况下，将相对迅速地蓄积突变（星号）。它将由于发生有害突变而成为一个无功能的假基因（ψA），后者将最终被从基因组中去除。然而，在某些情况下，突变差异将导致不同的表达模式或具有选择优势的其他性状（A2）。

复制性基因在复杂基因组中较为常见，并且它们能够在进化过程中迅速出现：一个普通基因大约平均每 1 亿年发生一次复制（Lynch and Conery，2000）。它们通过各种不同的机制而发生。某些将导致经常涉及单个基因的局部复制，并将在节 12.1.5 及 12.1.6 中讨论。其他则涉及大规模或整个基因组的复制，这将从染色体及基因组进化的角度讨论（节 12.2）。对其主要机制的介绍见框 12.3。

12.1.5 珠蛋白超基因家族通过一个基因复制、基因转化以及基因丢失/失活的过程进化而来

五类珠蛋白基因的进化

珠蛋白为含有卟啉的蛋白质，能够与氧发生可逆性结合，因而在呼吸系统中至关重要。它们可被分为 4 个功能组（表 12.1）：即已被彻底研究过的珠蛋白与肌红蛋白，以及较晚发现的脑红蛋白与胞红蛋白（Pesce *et al.*,2002）。珠蛋白在血液中运输氧气，属于由 16 号染色体上的 α 珠蛋白基因家族成员编码的两条完全相同的珠蛋白链，以及 11 号染色体上的 β 珠蛋白基因家族成员所编码的另外两条相同的链所组成的异源四聚体。肌红蛋白由 22 号染色体上的一个基因所编码。它属于单体蛋白，见于肌细胞中，在那里协助氧分子扩散至线粒体。14 号染色体上的脑红蛋白基因亦编码一个单体蛋白，表达于大脑中，并可增加脑组织的供氧（Burmester *et al.*,2000）。最近发现的胞红蛋白基因具有独特的表达，其功能尚属未知。

框 12.3 基因复制的机制与种内同源

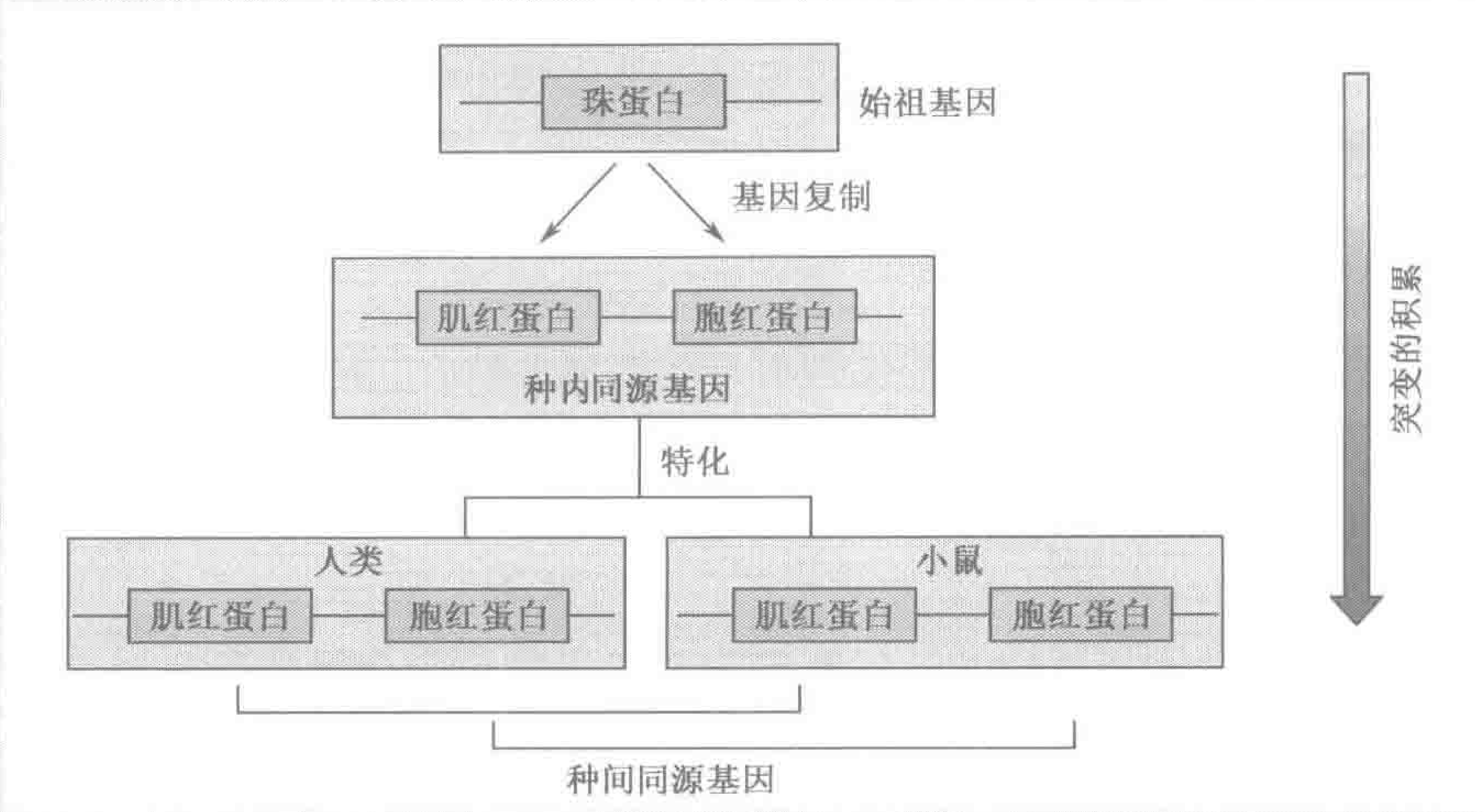
真核生物基因组中的基因复制意味着**同源基因**（homologous gene）［或称**同源体**（homolog）——具有显著的序列一致性的基因，提示相近的进化关系］可为两种类型之一（图）。**种间同源基因**（ortholog gene）为存在于不同的基因组中，因起源于一个共同的祖先而具有直接亲缘的基因。**种内同源基因**（paralog gene）为存在于同一个基因组中，由基因复制而产生的基因。各种不同的机制可造成基因复制。

- **基因串联复制**（tandem gene duplication）：单个基因可由于不等交叉或姐妹染色单体不等交换而发生串联复制（图 11.9）。成簇的人类基因家族通常起源于一系列的串联复制（例如 α 珠蛋白和 β 珠蛋白基因簇，见节 12.1.5）。相同的机制亦可导致大范围的重复，涉及含有若干基因的片段，但超大规模的串联复制相对少见，一定程度上是由于产生基因剂量效应的较大可能性。
- **反转座介导的基因复制**（retrotransposition gene duplication）（＝**复制性转座**，duplicative transposition）：这很可能是一种重要的因素。尽管通常将造成某个内含子基因的无内含子拷贝，它有时亦可能导致内含子序列的复制（节 12.1.6）。
- **基因横向（水平）转移**［horizontal(＝lateral) gene transfer］：这是指古老的基因在不同基因组之间的转移（Brown，2003）。国际人类基因测序协作组（2001）提出，数百个人类基因可能起源于脊椎动物种系进化中某些时间点上来自于细菌的横向基因转移。尽管这已被证明为不正确（框 12.4），但至少部分人类细胞核基因似乎起源于捕获原型线粒体基因组之后发生的古老的横向基因转移（节 12.2.1）。
- **节段性复制**（segmental duplication）：人类基因组的相当一部分是由密切相关的序列区段组成，后者分布于不同的基因组区域，在长度介于数千至数十万碱基之间的节段中呈现＞90％的序列

框 12.3 基因复制的机制与种内同源（续）

一致性。这类重复似乎发生在进化中很近的时期。见节 12.2.5。

▶ **多倍性**（polyploidy）：全基因组复制的有趣之处在于它立即为突变提供了基因拷贝，而不会带来由于基因剂量差异而产生的问题。若干物种的基因组曾明显发生多倍性，但这对人类基因的进化所造成的影响的程度仍有争议（节 12.2.3）。



同源体，种间同源基因，种内同源基因

肌红蛋白与胞红蛋白被认为起源于约 5 亿年前的基因复制。与其他珠蛋白基因的关系见图 12.4。

表 12.1 珠蛋白与人类珠蛋白基因的类型

蛋白质	功能	结构	多肽	基因	基因位置
珠蛋白* (Hb)	在血液中运输 O_2。Gower 1 与 Gower 2 为胚胎早期的类型。妊娠约 8 周时，胎儿的肝脏主要合成 HbF 及少量的 HbA。在新生儿中，70%的 Hb 为 HbF，但到一岁时，HbF 将降至 1%，而 HbA 将占优。HbA2 为见于儿童中的少量 Hb，在成人的 Hb 中占不到 3%	异源四聚体 Portland($\zeta_2\gamma_2$) Gower1($\zeta_2\varepsilon_2$) Gower2($\alpha_2\varepsilon_2$) HbF = fetal Hb($\alpha_2\gamma_2$) HbA 成人 Hb($\alpha_2\beta_2$) HbA2($\alpha_2\delta_2$)	α 珠蛋白 ζ 珠蛋白 β 珠蛋白 γ 珠蛋白 δ 珠蛋白 ε 珠蛋白	*HBA1*，*HBA2* *HBZ* *HBB* *HBG1*，*HBG2* *HBD* *HBE1*	16p13.3，α 珠蛋白基因簇 16p13.3，α 珠蛋白基因簇 11p15.5，β 珠蛋白基因簇 11p15.5，β 珠蛋白基因簇 11p15.5，β 珠蛋白基因簇 11p15.5，β 珠蛋白基因簇
肌红蛋白	在肌肉中运输及贮藏 O_2	单体	肌红蛋白	*MB*	22q13.1
脑红蛋白	主要表达于神经系统中，负责供氧，也可能为多功能性	单体	脑红蛋白	*NGB*	14q24
胞红蛋白	表达于几乎所有的组织中，具体功能未知	单体	胞红蛋白	*CYGB*	17q25.3

* **注**：由 α 珠蛋白基因簇（图 9.11）内 *HBQ1* 基因编码的 theta 珠蛋白多肽并不整合入 Hb 分子，其功能，如果有的话，尚属未知。

珠蛋白多肽之间的序列同源性提示该基因家族曾受到一系列基因复制事件的影响，某些较为古老，某些则相对较近。因此，来自不同染色体位置的珠蛋白多肽通常关系较远，提示较为古老的基因复制（虽然与肌红蛋白、脑红蛋白或胞红蛋白相比，α 与 β 珠蛋白基因家族之间的关系明显更近）。不同珠蛋白之间序列同源的程度，加上普遍保守的内含子位置（图 12.1），提示这四个基因家族起源于约 8 亿至 4.5 亿年前一个原始珠蛋白基因的一系列复制（图 12.4）。

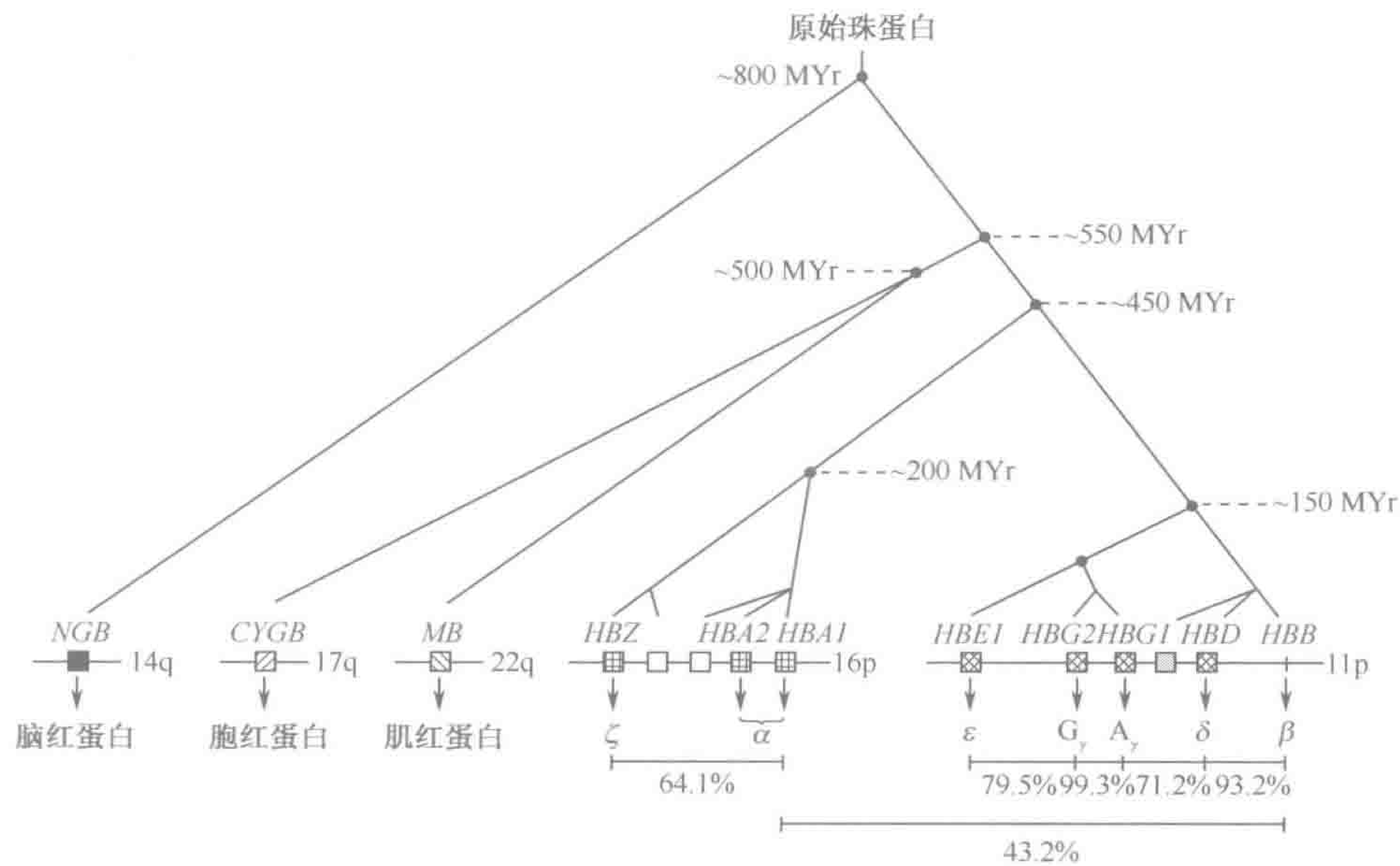


图 12.4　珠蛋白基因超家族的进化

由同一基因簇编码的珠蛋白较由分布于不同染色体上的编码产物具有更高程度的序列一致性（后者的序列一致性显著低于 30%，但 α 珠蛋白与 β 珠蛋白基因簇之间的比较例外）。α 与 β 珠蛋白基因簇后来通过串联基因复制而歧化。其中部分复制发生于很近的时期：*HBA1* 与 *HBA2* 基因编码完全一致的 α 珠蛋白；*HBG1* 与 *HBG2* 则编码相差仅一个氨基酸的 γ 珠蛋白。

随后，原始的 α 与 β 珠蛋白基因又发生了一系列的重复，其中一些发生在最近（图 12.4）。例如，人类的两个 α 珠蛋白基因 *HBA1* 与 *HBA2* 编码相同的产物，而两个 γ 珠蛋白基因 *HBG1* 与 *HBG2* 的产物相差仅一个氨基酸。在其他家族中，可能是因为有关的复制事件发生在相当长的时间之前，位于同一基因簇内的复制性基因具有更明显的序列分化。某些复制将产生常见的假基因。将来，人类的两个 α 珠蛋白基因以及两个 γ 珠蛋白基因之一或许会蜕变为假基因。

哺乳动物 β 珠蛋白基因簇的进化

对各种哺乳动物 β 珠蛋白基因簇的序列比较分析显示略有不同的基因组织结构。在不同种系中发生了单个基因的不同类型重复，另外，也存在基因丢失、融合以及基因转换（由供体基因复制而来的一段序列被用于置换受体基因的某些序列的非相互序列交换，其机制见图 11.10）的证据。例如，人类有一个 β 珠蛋白和一个 ε 珠蛋白基因，但

有两个 γ 基因；小鼠有两个 β 珠蛋白和两个 ε 珠蛋白基因，但却仅有一个 γ 珠蛋白基因（图 12.5）。在现代山羊的种系进化中，γ 珠蛋白基因发生了早期丢失，之后又发生了大规模的三倍重复（似乎是缘于两次连续的不等重组事件，其中前者造成了重复）。基因转化似乎相当频繁。例如，在人类和山羊的种系中，由 β 珠蛋白基因向 δ 珠蛋白基因的转化均很明显（图 12.6）。

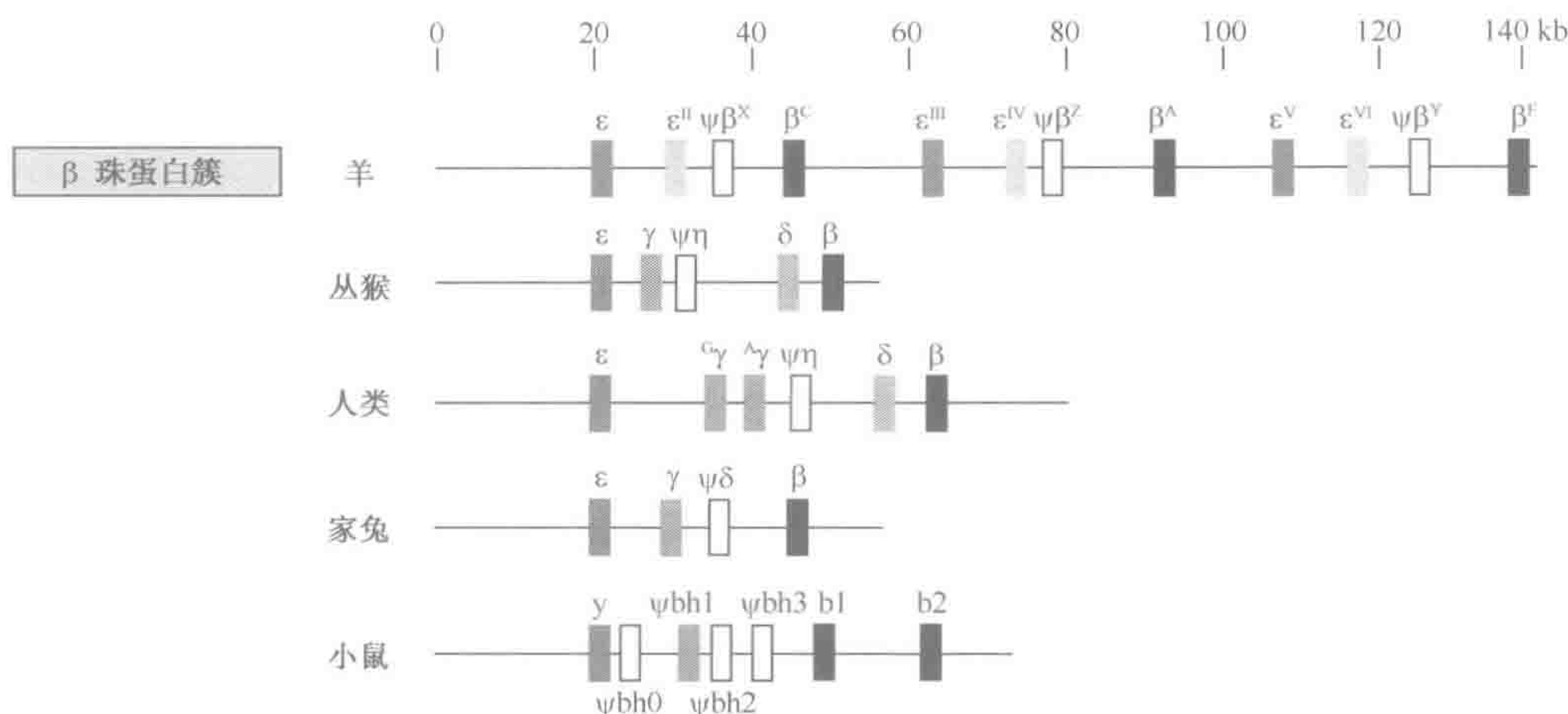


图 12.5　β 珠蛋白基因簇体现了种间同源的哺乳动物基因家族在结构上的显著差异

山羊 β 珠蛋白基因簇中的大量基因反映了一个串联的三倍重复事件。空白框代表假基因。据 Hardison 和 Miller (1993). Mol. Bio. Evo. 10，73～102. ©1993，由 Oxford University Press 重绘。

上述类型的事件均能够用该基因簇内的简单序列交换来解释。

由于诸如**负鼠**（opossum）之类的有袋类在该基因簇中仅有两个珠蛋白基因，即一个 β 珠蛋白基因与一个 ε 珠蛋白基因，现有的 β 珠蛋白基因簇可能起源于单一的珠蛋白基因，后者通过最初的复制产生了一个原型的类 β 以及一个原型的类 ε 珠蛋白基因。二者随后通过进一步复制而产生了该基因簇中的五种基因（图 12.6）。串联的重复序列之间最初完美的序列一致性使它们易于通过不等交叉（可引起进一步的基因复制、基因丢失以及基因融合）而发生进一步的序列交换和基因转化。然而，除非保守的选择压力能够维持它，否则经复制的序列终将分化，因此，尽管在长 1.6 kb 的个别基因序列之间存在相当的序列同源性，更长的旁侧序列之间则不然。

珠蛋白基因差异性表达的进化

串联型基因复制通常意味着调节序列与编码序列均被复制。随后发生的顺式作用调节区域的序列分化则可能导致表达在空间（例如在不同的组织中）与时间（例如在发育的不同阶段）上的改变。这可能是基因复制非常强大的一个进化优势。由于调节序列通常由极短的序列元件组成，突变可迅速地造成趋异的表达模式。当本质上就是同一个基因的表达拷贝具有各异的时空区间时，各拷贝将可能获得不同的功能：不同的环境将产生具有差异的选择压力，导致编码序列歧化，以最好的适应于不同的环境。

珠蛋白超基因家族提供了一些例子。一个原始的珠蛋白基因的古老复制产生了拷

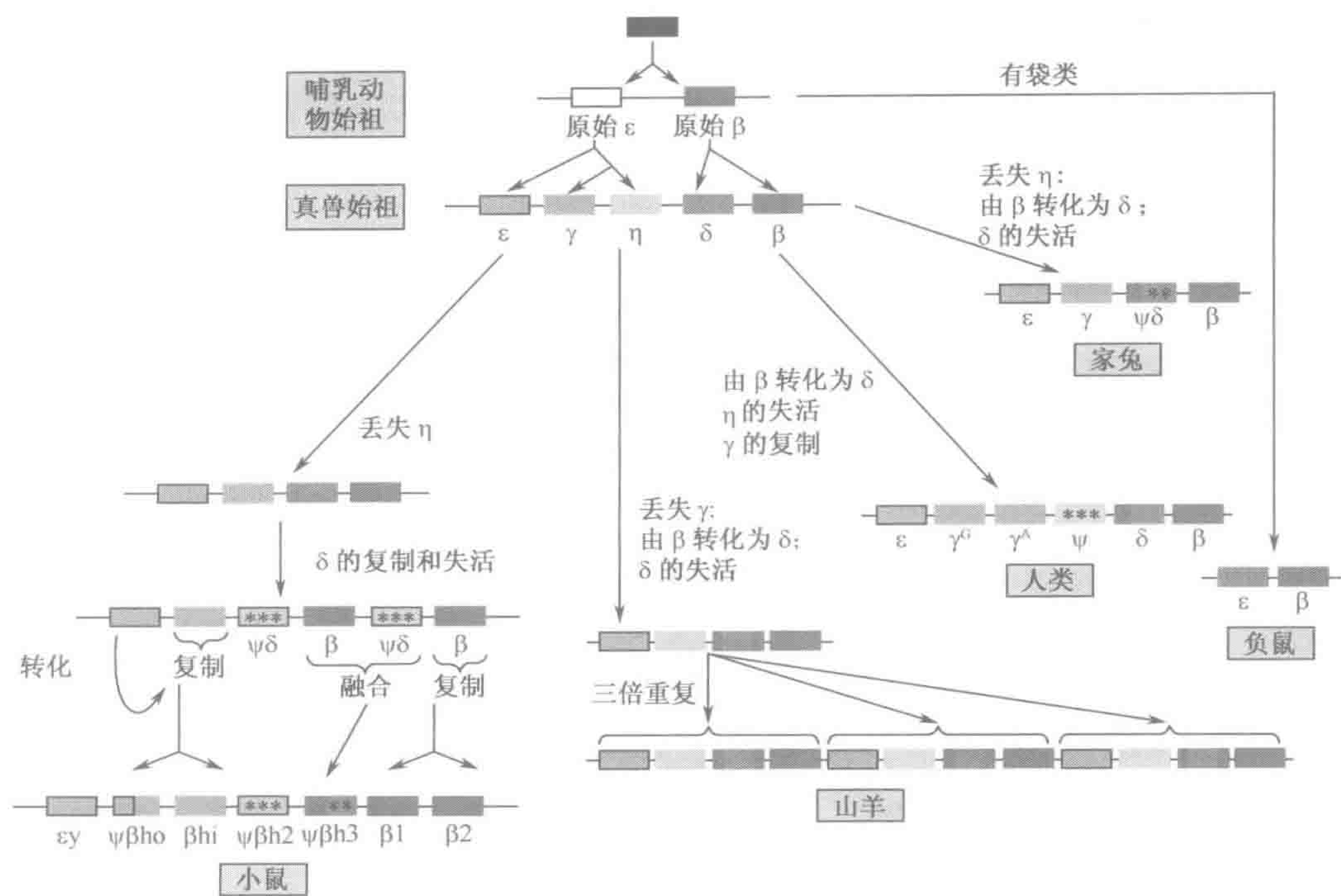


图 12.6　哺乳动物β珠蛋白基因簇的进化曾涉及频繁的基因复制、转化以及基因丢失或失活

值得注意的是，除基因复制和基因丢失外，亦存在常见的某一基因的序列显示证据拷贝自另一基因的例子。这在本图中被松散地描述为转化，但在某些情况下却可能涉及基因转化以外的机制。例如，在兔子的种系进化中，由β基因向δ基因的转化可能涉及一次与小鼠种系进化中 *ψh3* 基因的产生过程类似的不等交叉。据 Tagle 等（1992）. Genomics 13，741～760 重绘，经 Elsevier 出版社允许。

贝，后者的调节序列所发生的歧化又造成了它们在不同组织（血液、肌肉、神经系统等）中的表达。不同的拷贝对于环境的适应导致了其产物在序列上的显著不同，分别形成了血红蛋白、肌红蛋白以及脑红蛋白等——每种仍具有与氧结合的核心功能。组成血红蛋白的珠蛋白随后的精细化产生了不同种类的珠蛋白，后者逐渐表达于特定的发育阶段并似乎变得更加专一。譬如，ε、ζ 以及 γ 珠蛋白链似乎特别善于在早期发育相对缺氧的环境中与氧结合，而 α 与 β 珠蛋白链则可能是成体组织环境中最理想的多肽链。

12.1.6　反转座可能容许外显子混编并且是基因进化的一个重要原因

反转座，即由 RNA 转录物所产生的天然 cDNA 拷贝被插入到一个新的染色体位置，是塑造人类基因组的一个非常有力的工具。超过 40%的人类基因组由衍生自反转座的重复组成，这些重复中的一小部分仍在进行活跃的转座（节 9.5.1）。反转座机制通过制造外显子的拷贝并将其从一个基因组位置穿梭到另一个位置，以及为一些基因复制提供条件从而在基因的进化中起到重要的作用。

由 LINE 介导的外显子混编

LINE1(L1) 元件属于非 LTR 类型的反转座子，能够自行转座（节 9.5.2）。在实

验条件下，LINE1 元件显示能够插入到一个基因的内含子中，对其下游外显子进行复制，之后再将该拷贝转座到另一基因中（Moran *et al.*,1999）。这是因为 LINE1 反转座机制对自身的 3′端仅有微弱的特异性，导致其对该末端调节信号的忽略。在某个 LINE1 元件插入一个基因之后，对于 LINE1 重复的转录通常将绕过其自身的多聚腺苷酸［Poly(A)］序列而使用其宿主基因位于下游的多聚腺苷酸信号。通过这种方式，该元件可复制宿主的某个外显子，后者将在下一次反转座［**LINE 介导的 3′端转导**(LINE-mediated 3′transduction)，图 12.7］时被拼入另一个基因中。这一机制因此可能是似乎在基因进化中扮演重要角色的外显子混编的基础。

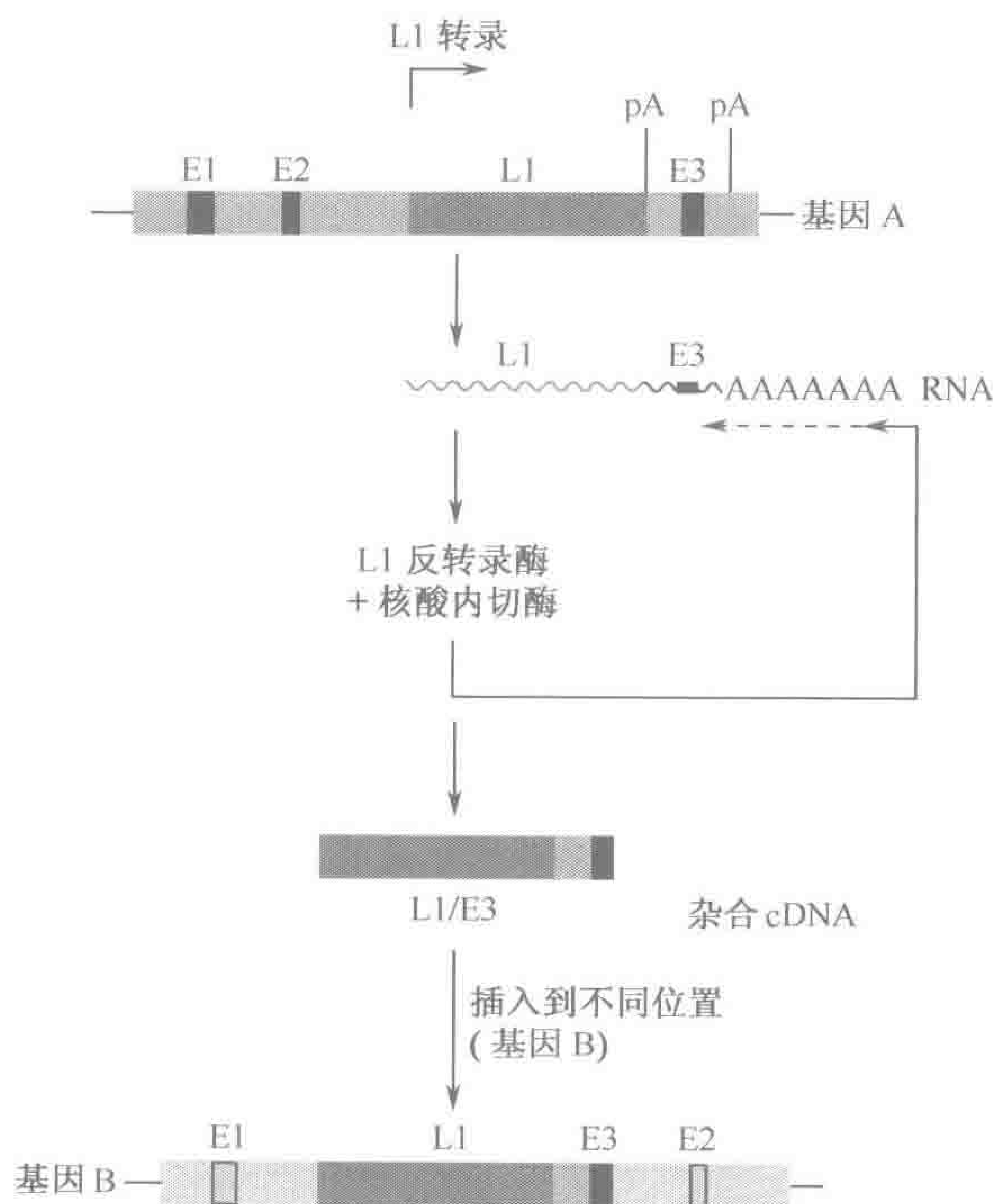


图 12.7 基因之间的外显子混编可能由可转座元件介导

LINE(L1) 序列家族含有能够在人类基因组中进行活跃转座的成员。LINE 元件具有弱的多聚腺苷酸［Poly(A)］信号，因此转录可越过这一信号直至位于附近的下一个 Poly(A) 信号，如图上方的基因 A 所示。产生的 RNA 拷贝因此不但含有 LINE1 序列，而且也含有一个下游外显子（如本例中的 E3）。LINE1 反转录酶复合体随后将作用于延伸的 Poly(A) 序列，产生一个同时含有 LINE1 与 E3 序列的 cDNA 拷贝。随后向一个新的染色体位点的转座将导致外显子 3 被插入到一个不同的基因（基因 B）中（Moran *et al.*,1999）。

通过反转座来进行的基因复制

基因序列的反转座是在复杂基因组形成中的常见事件。对源于一个含有内含子基因经剪接的 RNA 的复制意味着含有内含子的序列已被去除。当被复制的 RNA 是 mRNA 时，复制产物中将没有任何启动子序列。因此，对 mRNA 序列复制物的反转座通常将产生无活性的已加工的假基因（图 9.14）。然而，这类 cDNA 拷贝偶尔也会被整合到一个功能性启动子附近，并可能会受到选择压力以保持其基因功能，从而成为一个**反基因**

(retrogene)。经典的例子就是由常染色体加工后的、功能上为精子发生所必需的 X 连锁基因。在精子发生的过程中，X 与 Y 染色体浓缩而形成无转录活性的**XY 小体**（XY body），而一个常染色体的基因拷贝则可以提供必需的基因产物。衍生自非 X 连锁基因的反基因亦已被发现（节 9.3.6；表 9.11）。

直到最近，通过反转座进行基因复制被认为仅限于拷贝外显子的序列。然而，对于一个镶嵌型新基因的分析（Courseaux and Nahon，2001）已证明内含子序列在某些情况下亦可能由间接途径被复制。这可能发生在一个含有内含子基因的反义链产生一条成熟的 RNA 转录物，而后者中含有与转录自另一条链的基因的内含子互补的序列时。由于反义的 RNA 转录物已知在人类基因组中相当常见，反转座的拷贝可能有时会导致含有内含子基因的复制。

12.2　染色体与基因组的进化

12.2.1　线粒体基因组可能起源于一个原核细胞被某个前体真核细胞内吞之后

除细胞核之外，线粒体也具有一个基因组，植物细胞中的叶绿体亦然。线粒体与叶绿体基因组的结构及表达与原核细胞具有显著的相似性（见下图），提示真核细胞起源于某个原始真核细胞（**原型真核细胞**，protoeukaryote）吞入了某种类型的原核细胞（**共生生物**，symbiont）之后。这一过程被认为给产生的新细胞带来了选择上的优势，并被称为**内共生**（endosymbiosis）（图 12.8）。

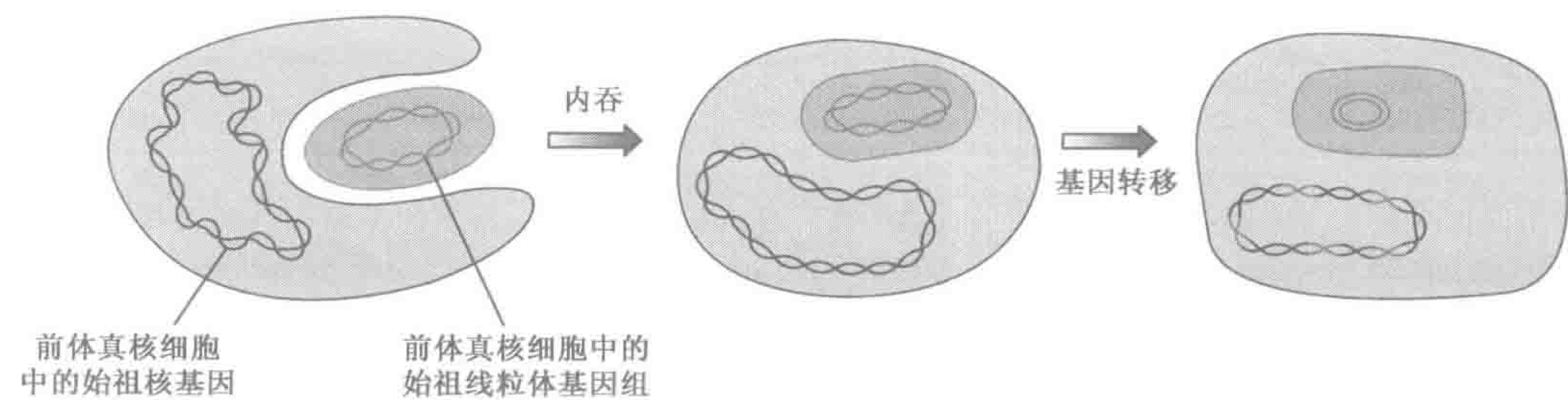


图 12.8　人类线粒体基因组可能起源于某个原始真核细胞将一个原核细胞内吞之后
在这个特例中，正在进行内吞的细胞没有细胞核，但在一些模型中则被想像为一个有核的原始真核细胞所进行的内吞。内吞之后，原核细胞基因组中的基因被推测转移到了前体的核基因组中，剩下一个显著减小的线粒体基因组。基因转移的一种可能机制见 Doolittle(1998)。

内共生生物假说推测被吞入的原核细胞的基因组产生了今天的线粒体基因组。现代的原核细胞（细菌及藻类）通常含有一至数百万对 DNA 碱基以及数百至几千个基因，但线粒体基因组却小得多。例如，人类线粒体基因组长约 16 kb，仅含有 37 个基因，而且绝大多数线粒体蛋白质及相关功能均由细胞核基因所编码（节 9.1.2）。因此，最初存在于被吞入细胞内的许多基因可能已通过**横向或水平基因转移**（horizontal or lateral gene transfer）进入宿主细胞的基因组（Doolittle，1998）。在细胞进化的早期，横向基因转移可能相当广泛。

对于涉及内吞而产生线粒体的细胞的性质仍存在争议。最初的假说曾推测宿主细胞

具有某些真核细胞的特征，但更近的假说却认为宿主是一个藻类细胞。**氢假说**（hydrogen hypothesis）提出真核细胞起源于一种内吞了某种产氢的 α 蛋白菌、厌氧且具有氢依赖性的藻类宿主细胞。另一种**互养共栖假说**（syntrophic hypothesis）则推测产氢的共生细菌为一种 δ 蛋白菌。上述类藻宿主被推想为严格的自养型（能够将环境中的简单分子加工成自身的有机物），而不是异养型，即依赖于摄入由其他生物所制造的有机物。然而在后来，为了省去在其细胞质中毫无意义的代谢循环，宿主细胞丢失了其自养通路，一个含有原始线粒体的不可逆的异养型细胞从此诞生，但不再依赖于氢。许多此类生物开始使用效率更高的氧基呼吸，需氧的线粒体因而得到进化。

氢/共营养假说被建立在对某些缺乏线粒体的真核生物，如贾第氏虫（*Giardia*）、副基体类如毛滴虫（*Trichomonas*）、阿米巴原虫，以及某些纤毛虫及真菌等的观察之上。被广泛接受的系统发生分类法将贾第氏虫与毛滴虫视作最早从真核生物种系中分化出来者（框 12.4；图 12.22）。副基体类，以及某些缺乏线粒体的纤毛虫与真菌，具有被称为**氢小体**（hydrogenosome）的细胞器，后者可发酵丙酮酸并产生氢。氢小体并不含有任何有助于追溯其起源的基因组，然而，针对氢小体靶蛋白的系统发生分析提示其可能与线粒体具有共同的进化途径，并且氢小体本身即可能是一种高度分化的线粒体。缺乏线粒体的真核生物亦具有类似于细菌的代谢酶（除其他与细菌相似的基因之外）。似乎同等重要的是，大多数真核生物利用组蛋白包装其核 DNA，而唯一具有组蛋白与核小体的原核生物是广古菌（*Euryarchaeota*），即包括耗氢的甲烷生产者在内的一个藻类分支。进一步的细节与其他假说见 Brown（2003）。

框 12.4　统一的生命之树与横向基因转移

分子系统发生学（Molecular phylogenetics，根据蛋白质或核酸之间的亲缘关系对有机体进行分类）

在二十世纪七十年代后期造成了巨大的震撼：rRNA 分析显示，一组产甲烷细菌显著区别于其他细菌。正如它们最初的称呼那样，**藻菌**（archaebacteria）似乎亦具有一些非同寻常的细胞特征。在信息传递的许多方面（DNA 复制、DNA 修复、转录、翻译等），它们表现出与真核细胞的密切联系。因此，越来越多的人同意，过去将生命划分为原核与真核生物的分类法应该被替换成三个**域**（domain）。

- 细菌——常见的原核生物，在过去已被详细地研究过（如革兰氏阴性与阳性菌、蓝细菌等）
- 藻类——原核生物，但具有与真核生物相似的信息传递过程。通常被分离于极端的环境中（如温泉、高盐环境、极端的 pH 值等），但亦存在于更普通的场合，包括土壤与湖泊中，它们亦被发现兴旺于牛的消化道、白蚁以及海洋生物中，并在那里产生甲烷。
- 真核生物

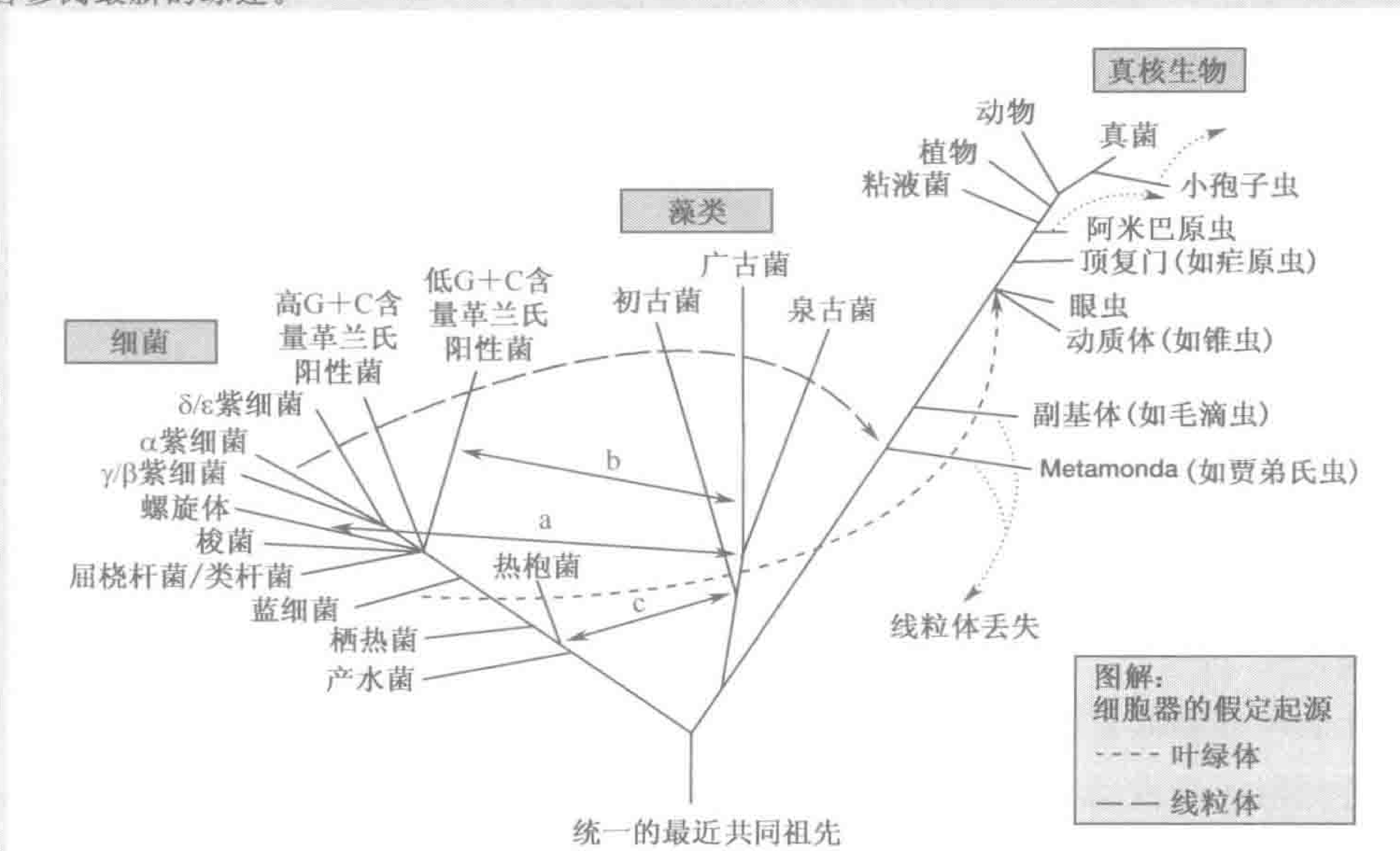
其中，细菌被推想为率先从全体的最近共同祖先中分离出来（见图）。线粒体与叶绿体被认为起源于原始真核细胞对于某种原核细胞的内吞，之后被内吞者基因组中的许多基因转移到宿主的基因组中，即**横向基因转移**（horizontal gene transfer，HGT）的一种形式。

横向基因转移似乎曾在基因组的进化中扮演关键角色，但若干较近的自然发生的横向基因转移的例子已被发现，尤其是在不同种类的细菌之间（如质粒与噬菌体所介导的致病及抗生素耐药基因的转移）。真核生物基因组之间亦可能发生横向基因转移（例如在不同种的果蝇之间通过 DNA 转座子进行）。国际人类基因组测序协作组（2001）亦得出惊人的结论，即数百个人类基因起源于脊椎动物种系进化的某些点上来自细菌的横向基因转移。这些人类基因呈现明显的与细菌基因的同

框 12.4　统一的生命之树与横向基因转移（续）

源性，却似乎并不存在于某些无脊椎动物（果蝇、秀丽新小杆线虫）的基因组中。然而，目前已知这一结论并不正确，因为进一步的系统发生分析已在其他无脊椎动物中发现了同源体，因此人类的基因可从具有共同祖先的角度来解释——一些物种中该同源体的缺乏可以被解释为某些种系进化中的基因丢失（Brown，2003）。

本图所示意的共同生命之树远非已被普遍接受，对于其他各种蛋白质序列数据的研究已经对其提出了质疑。不可避免的是，如果必须将常见的横向基因转移考虑在内的话，系统发生学分类将会变得困难，因为不同的基因集合将可能产生冲突性的系统发生学结果。因此，**从全基因组的角度构建进化树**（whole genome approaches to tree construction）正在进行。这一领域进展很快，建议读者参阅最新的综述。



基于 rRNA 系统发生的共同生命之树

除涉及线粒体与叶绿体基因组的形成之外，横向基因转移亦被发现于其他场合，如螺旋体与藻类之间（箭头 a）；低 G+C 含量革兰氏阳性菌与藻类之间（箭头 b）以及嗜热菌与藻类之间（箭头 c）等。经 Nature Publishing Group 允许，复制于 Brown（2003）. Nature Rev. Genet. 4，121～132。

12.2.2　减小的选择压力导致了线粒体密码子的歧化

人类线粒体的遗传密码与细胞基因组以及植物线粒体基因组所使用的‘通用’遗传密码略有不同（图 1.22）。尽管与其他哺乳动物线粒体基因组的遗传密码完全一致，但它也呈现某些与其他真核生物如果蝇及真菌的线粒体所具有的非通用密码的不同之处。

线粒体遗传密码被认为曾因选择压力的减小而发生过分化。产生线粒体基因组的被内吞原核生物的基因组最初可能含有了至少数百乃至于上千个基因。与所有的大型基因组相似，它可能曾受到强大的保守性选择压力以维持其通用的遗传密码（密码的轻微变化即可能导致基因功能的缺乏）。

随后，也许是通过细胞器溶解、DNA 整合入核基因组中、细胞器原拷贝丢失以及

基因漂移对之进行固定等相继过程，源自原始线粒体基因组的基因被转移至核基因组中(Doolittle，1998)。由于向细胞核基因组进行的基因转移将不断降低基因的编码潜力，保留原始遗传编码的选择压力也可能越来越小。

最终，一个严重减员的基因组终于形成（人类线粒体基因组中仅含有 13 个编码多肽的基因）。由于仅涉及极少数多肽，维持通用遗传密码的选择压力可能已变得松懈，正常密码的翻译在一定程度上的漂移将被容许而不会引起灾难性的后果。另一种可能的情况是，变化了的密码（图 1.22）并未被用于氨基酸替换将造成有害影响的部位。

当然，线粒体基因组耗损的过程是一个贯穿后生动物整个进化的漫长过程。其结果就是，不同物种在线粒体基因组所保留的基因上呈现出差异（因此一些物种具有不同的线粒体遗传密码）。然而，植物的线粒体基因组保留了相对较多的基因，因而不可能发生通用遗传密码的漂移。

12.2.3　脊椎动物基因组的进化可能曾涉及整个基因组的复制

导致**多倍体**（polyploidy）的**基因组复制**（genome duplication）是增加基因组长度以及多样性的有效途径。由于所有基因同时获得了拷贝，避免了出现基因剂量差异的问题。一些物种，包括大多数开花植物、多种鱼类、酿酒酵母（*Saccharomyces cerevisiae*）、爪蟾（*Xenopus laevis*）以及至少一种哺乳动物，即四倍体南美红兔鼠（Otto and Whitton，2000；Wolfe，2001）等，均为自然存在或已蜕变的多倍体。组成性四倍体在人类中非常罕见，且为致死性，但人类与其他二倍体生物均具有一些因发生未间插细胞分裂的连续有丝分裂所造成的染色体复制或者细胞融合而产生的自然状态下的多倍体细胞（节 3.1.4）。

由于若干物种明显曾发生过基因组复制，多倍体化的产生事件显然并非罕见。这一现象，连同对于脊椎动物与无脊椎动物基因的特征的比较，提示所有脊椎动物在进化中可能均发生过至少一轮基因组复制。在基因组复制以及短暂的四倍体状态之后，大规模的染色体重组估计可能造成了染色体的分化及二倍体的恢复，但染色体的数目则已经加倍。

在基因组复制之后对于许多基因拷贝的选择压力的松懈将导致其中许多发生有害突变而成为假基因。随后，由于不存在任何保守性的选择压力，缺陷的基因拷贝将可能通过各种 DNA 更新的机制从基因组中被清除（Lynch and Conery，2000）。因此，在基因组复制相当长时间之后，既往的基因组范围复制的证据将变得稀少：仅有很少功能活性得到保存的复制性基因会被保留下来。

最初有关在脊椎动物的进化中所发生的古老的四倍体化事件的提议推测曾发生过两轮基因组复制，即 2R 假说（Wolfe，2001）。支持 2R 假说的许多证据均建立在对在无脊椎动物如果蝇等中具有关键功能的基因及基因簇在脊椎动物中均具有 3～4 份对应物的观察之上。重要的例子包括 4 个经典的 *Hox* 基因簇（图 12.9）、4 个 *paraHox* 基因簇（人类染色体 13q13-14，Xq13-q22、4q11-12、5q31-33 区域）以及 MHC 基因簇（人类染色体 1q21-q25、6p21.3-p22.2、9q33-q34、19p13.1-13.4 区域；Abi-Rached *et al.*，2002）。上述每种在与脊椎动物关系最近的非脊椎动物文昌鱼（*Amphioxus*）中均仅有一个对应的基因簇。

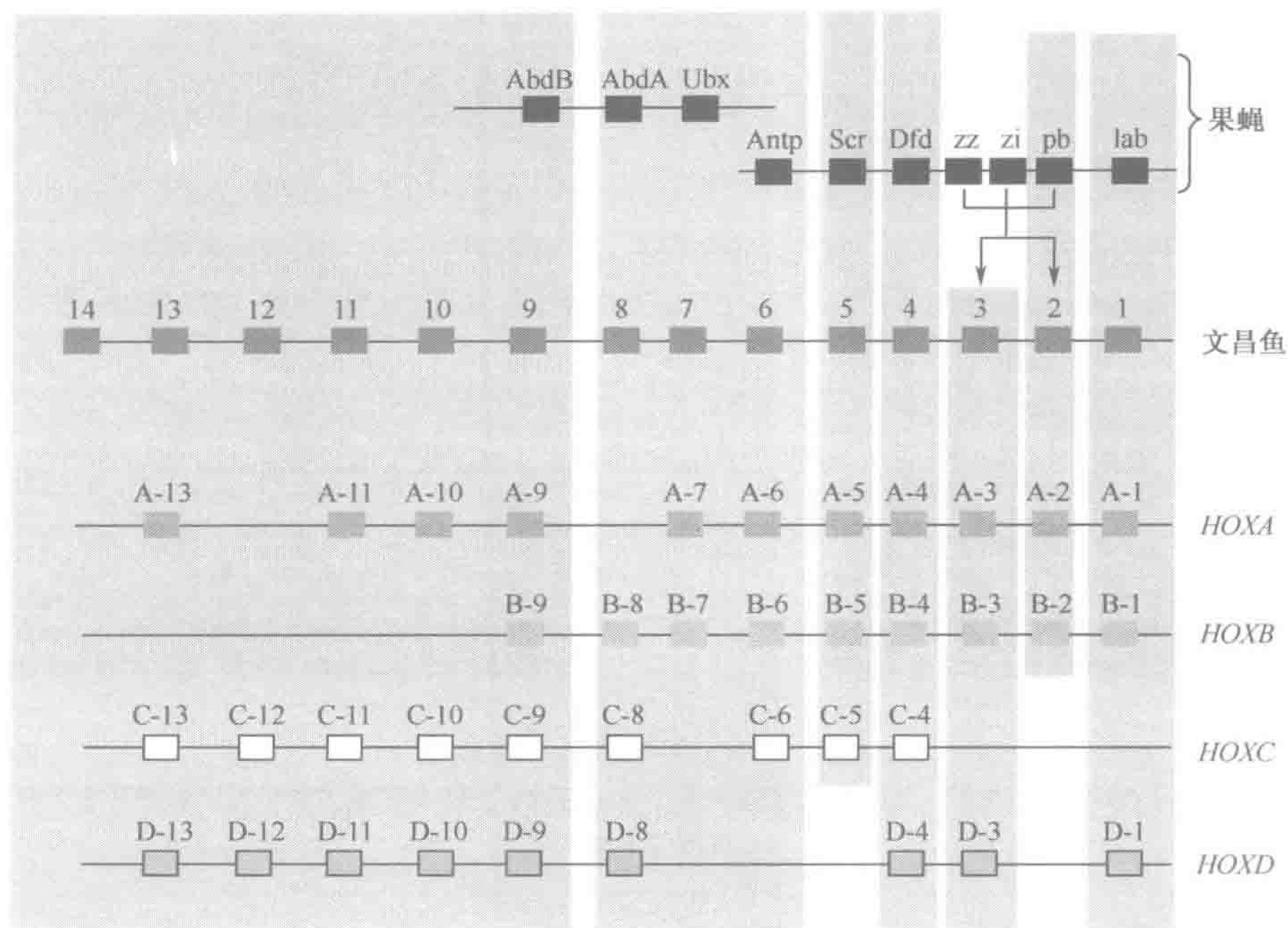


图 12.9　哺乳动物与文昌鱼 *Hox* 基因簇的结构表明曾发生过一或两轮原始的基因组复制

人类以及其他哺乳动物具有 4 个含有 9～11 条典型 *Hox* 基因的基因簇，而被认为与脊椎动物关系最近的文昌鱼则具有 14 条这样的基因。基因簇中基因的线性顺序被认为可能决定其在发育中的表达时序及沿身体前——后轴表达范围的前部边界（图 3.10）。涂黑的方块表示由具有非常相似的表达模式且功能可能相似的基因组成的种内同源基因群。虽然 4 个经典的哺乳动物 *Hox* 基因簇均呈现对于基因顺序的普遍强烈保守，种内同源群则提示含有 13 或 14 条基因的原始基因簇可能曾发生基因丢失。发生于果蝇种系中的一次重组导致上述基因被分隔成两个亚簇。

在基因组计划之后，变得清晰的是脊椎动物具有大约30 000～35 000个基因，近乎发现于无脊椎动物中基因数目的两倍、而非广泛推测的四倍。由 McLysaght 等（2002）最近对人类基因组草图所作的一个全面分析提示后者具有较随机发生多得多的复制性基因。约 25%的人类基因具有明显相关的种内同源体，与果蝇以及秀丽新小杆线虫种间同源基因的比较提示在大约 3.5 至 6 亿年前基因复制曾大量发生，这与至少曾发生过一轮全基因组复制的推测一致。

12.2.4　在哺乳动物基因组的进化中曾发生无数次大型染色体重组

在哺乳动物基因组的进化中，大规模的染色体重组曾相当频繁，而且染色体的进化可以独立于表型的进化之外。一个经典的例子就是两种麂（一种小型的鹿）。中国麂与印度麂（图 12.10）的血缘是如此相近，以至于二者能够交配产生后代，然而后者并不能存活下来。中国麂有 46 条染色体，而由于各种染色体的融合事件，印度雌麂与雄麂仅分别有 6 条及 7 条染色体。

虽然上述的例子属于非常特殊的一个，但在哺乳动物的进化中，显著的染色体重组却在不断地发生。倒置似乎尤其频繁，易位则相对较少。着丝粒（由迅速演化的序列构



图 12.10　中国麂与印度麂

中国麂（*Muntiacus reevesi*，左图）与印度麂（*Muntiacus muntjak*，右图）亲缘关系很近，但却具有大不相同的核型（正文）

成）可改变位置。将人类染色体同与我们亲缘最近的现存物种、即大型猿类相比，二者的染色体显带模式非常强烈地相似（Yunis and Prakash，1982）。最为常见的重组为倒置（包括臂间及臂内倒置），此外还有一些易位（节 12.4.2 及图 12.27，图 12.28）。

在亲缘很近的物种中发现种间同源染色体较为简单，而将亲缘较远物种的染色体进行比较则通常显示仅较小的片段得到了保留。线性基因顺序的保守（**同线保守**，conservation of synteny）因此常常局限于较小的染色体片段。为了评估同线保守，既往的办法是将种间同源性基因定位于两个物种的染色体上，而基因组计划则提供了更为详尽的同线保守性图。人-小鼠的同线保守性见图 12.11。

人与小鼠所共有的约 342 个染色体片段意味着同线保守性平均延伸了不到 10 Mb，而人与小鼠的大多数染色体序列在其他各种物种的不同染色体上均有种间同源体。X 染色体则是引人瞩目的一个例外，人类 X 染色体上的绝大多数序列在小鼠 X 染色体上均有种间同源体（因为哺乳动物的 X 染色体失活已演化成为常染色体：X 染色体连锁基因之间事实上 2：1 的基因剂量比例的保证，节 12.2.8）。然而，即使对于 X 染色体来讲，亦发生过无数次倒置，搅乱了人类与小鼠基因的顺序，对于常染色体上保守性同线片段长度与数目的分析结果则与随机发生的染色体断裂过程相一致（小鼠基因组测序协作组，2002）。

12.2.5　灵长类种系进化中的节段性复制以及着丝粒周边与亚端粒序列在进化上的不稳定性

对人类基因组进行测序所带来的一个大的惊奇就是在最近的进化中所发生的**节段性复制**（segmental duplication），即一种可导致染色体重组的重要形式的亚基因组范围复制的广泛存在。约 5%的人类基因组序列目前被认为属于散在的复制拷贝，其中序列一

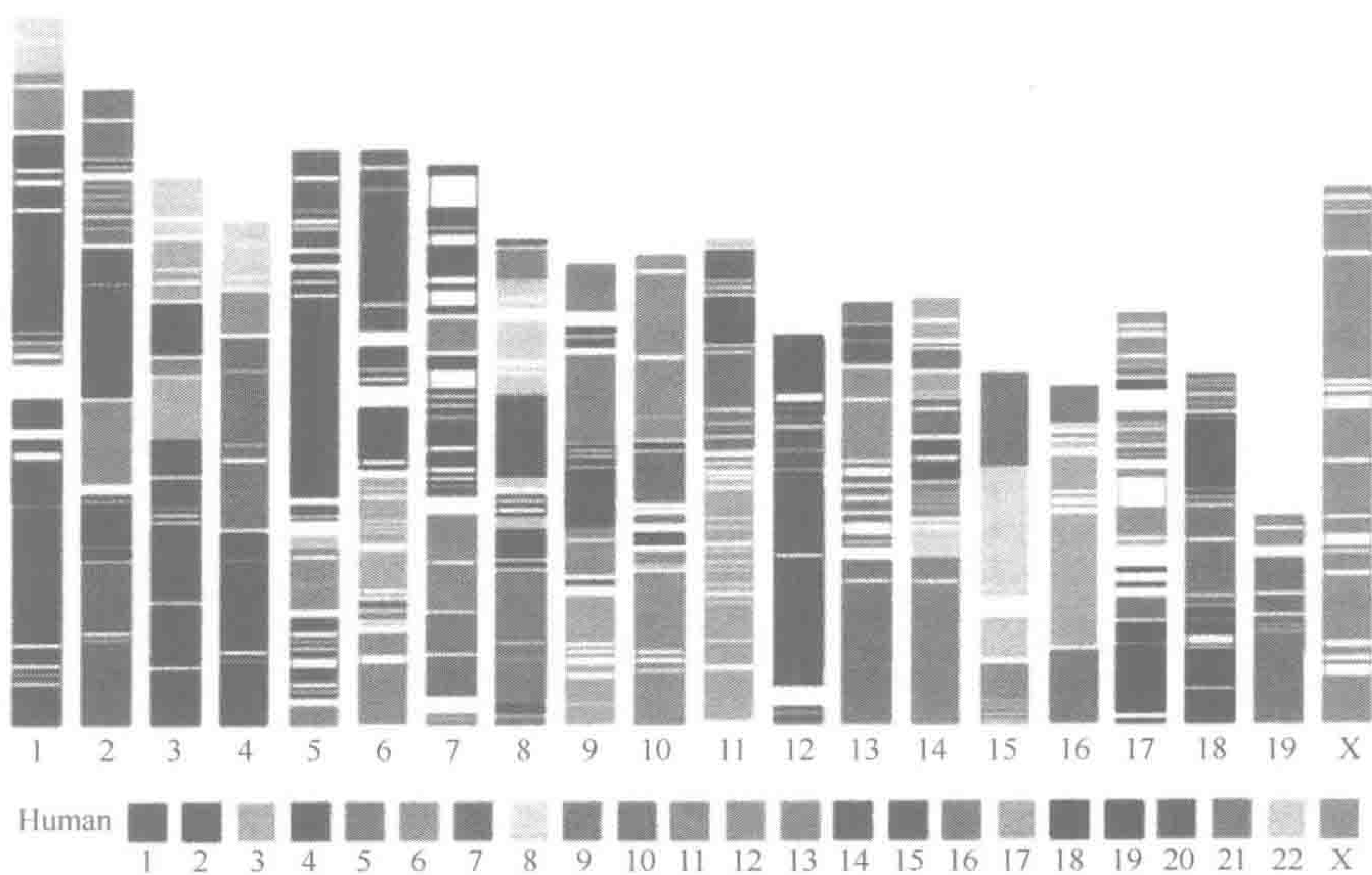


图 12.11　人——小鼠同线保守通常局限于较小的染色体片段

人类染色体上长度>300 kb 的保守同线片段与区块被重叠在 20 条小鼠染色体（上方）上。每种颜色对应于一条特定的人类染色体。图中的 342 个片段由细白线分隔为 217 个同色的区块。X 染色体由单一、交互的同线区块表示，而人类 17 号和 20 号染色体则完全对应于小鼠 11 号和 2 号染色体的一部分（但前者已通过剧烈重组变成至少 16 个片段）。其他染色体则呈现广泛得多的染色体内重排。经 Nature publishing group 允许，复制于 Mouse Genome Sequencing Consortium (2002). Nature 420，520～562。

致性>90%的区段长度从 1 kb 延伸至数百 kb(Bailey *et al*.,2002；Samonte and Eichler，2002；图 12.12)。它们包括：

- **染色体内复制**（intrachromosomal duplication）倾向于发生在常染色质区。当序列一致性超过 95%且长度为 10 kb 以上时，这些序列元件将经常诱发大规模的缺失、重复及倒置，其中许多与疾病相关（节 11.5.4，11.5.5）。
- **染色体间复制**（interchromosomal duplication）倾向于发生在着丝粒周边或亚端粒区。这类复制可能涉及基因或基因的一部分，后者将产生散在的无功能基因拷贝(图 9.13)、镶嵌型转录物以及潜在的新基因等。在灵长类中，染色体间复制似乎略少于染色体内复制（Samonte and Eichler，2002）。

在小鼠基因组中筛查与节段性复制的人类序列密切相关的序列通常仅能发现单一拷贝的序列。人类的节段性复制因此起源于相对最近的进化阶段：节段性复制事件似乎在过去四千万年左右的灵长类种系进化中一直发生着，其中大部分可能发生于过去的一千二百年中（Samonte and Eichler，2002）。然而，节段性复制并不局限于灵长类：近期独立发生的着丝粒周边的复制曾发生于小鼠中（Thomas *et al*.,2003）。节段性复制的机制尽管尚不清楚，但通常可形成镶嵌性结构，导致来自不同染色体部位的序列模块之间的组合。基因组中的特定区域，尤其是着丝粒周边似乎易于接受来自其他基因组区域的序列拷贝，并将它们与其他类似的区域进行交换（图 12.13）。

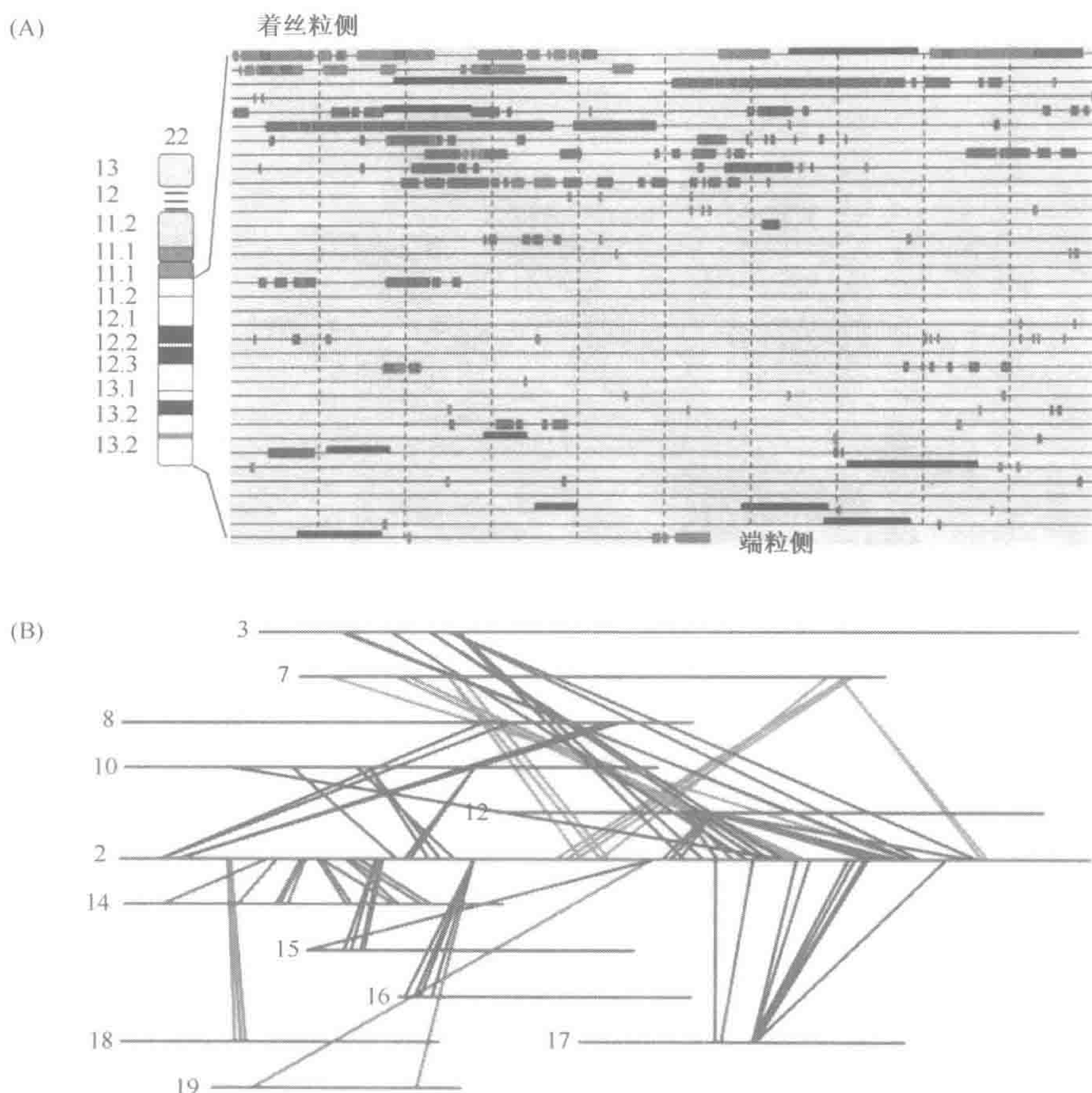


图 12.12 人类基因组中近期发生的节段性复制举例

(A) 22q **上的节段性复制** 水平线条表示着丝粒端（左上方）与端粒（右下方）之间连续相隔 1 Mb 的序列。黑色条块代表序列之间的缝隙。涉及染色体间复制的序列（红色）多数局限于最靠近着丝粒和端粒的区域。染色体内复制由蓝色表示。经 Elsevier 出版社允许，引自 Eichler(2001). Trends Genet. 11，661～669。(B) **人类 2 号染色体序列的染色体间同源性**。2 号染色体由图中部的红色横条表示。11 条其他染色体（绿色横条）含有与 2 号染色体高度同源的序列（由彩色竖线连接）。经 American Association for the Advancement of Science 允许，复制于 Venter 等（2001). Science 291，1304～1351。

12.2.6 人类的 X 和 Y 染色体呈现包括常见假常染色体区域在内相当范围的序列同源性

在哺乳动物中，成对的同源性常染色体在结构上基本一致（同态的）；发生于减数分裂期的染色体配对被推测借助于同源体之间高度的序列一致性，尽管其机制尚属未知。与此形成对照的是，人类与其他哺乳类物种的 X 与 Y 染色体为异态的。人类的 X 染色体属于近中着丝粒染色体，含有超过 160 Mb 的 DNA，而 Y 染色体则属于近端着丝粒染色体，并且要短得多（含有约 50 Mb DNA)。人类的 X 染色体含有众多的重要基因，而与此形成鲜明对照的是，Y 染色体仅含有约 50 个基因（表 12.2），大体上由无遗传活性的组成性异染色质构成。然而，值得注意的是，位于 Y 染色体非重组部分

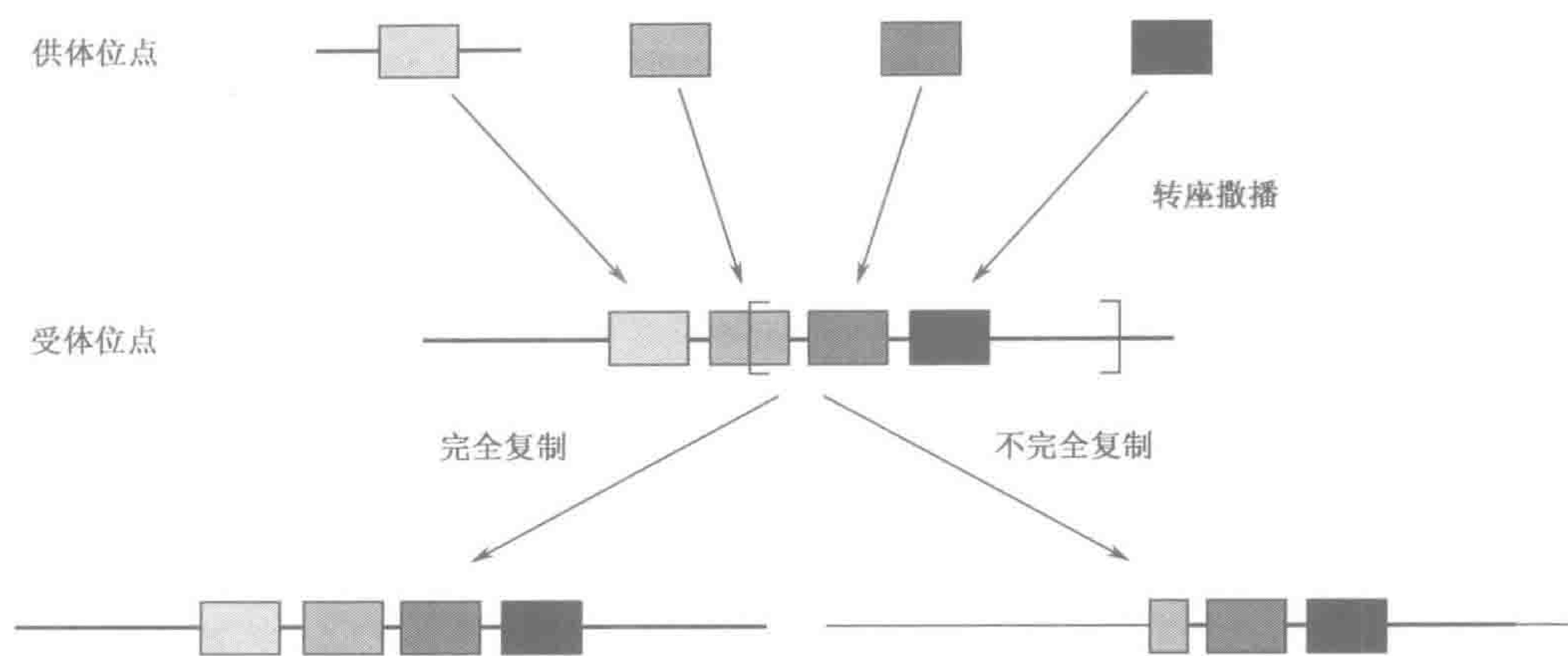


图 12.13　节段性复制的模型

中等至较长的序列拷贝（1～200 kb 的节段）起源于对基因组中特定区域（供体位点）的复制并转移、整合入基因组的其他部位（复制性转座）。如图所示，源自不同区域的供体序列可插入相同的受体区域。不同的复制性转座可能独立发生于不同的时期，产生较大的具有镶嵌结构的重复序列区块。这一镶嵌结构的局部可能接着被复制并拷贝到其他基因组区域。随后的重组（缺失及倒置）将改变这些区域的结构。经 Nature 出版集团允许，复制于 Samonte and Eichler(2002). Nature Rev. Genet. 3，65～72。

的许多基因均与精子发生/性别决定有关（表 12.2），而 X 染色体亦似乎富含性别与生殖相关基因（Wang *et al*.,2001）。

表 12.2　人类 Y 染色体基因的功能分类

基因种类	基因	已知/假定功能	表达特异性	Y 染色体上的多个拷贝？	具有活性的 X 染色体同源体？	X 同源体在女性中失活？
假常染色体区	Many*	同常染色体基因一样多样	多样	否	是	是（*SYBL1*、*HSPRY* 除外）
NRY 1 类	*RPSRY*,*ZFY*,*USP9Y*,*DBY*,*UTY*,*TB4Y*,*SMCY*,*EIF1AY*	持家基因	广泛	否	是	否
NRY 2 类	*TTY1*,*TSPY* *PRY*,*TTY2*,*CDY*,*XKRY*,*DAZ*,*BPY2*	精子发生	睾丸	是	否	不适用
NRY 3 类	*SRY*	男性性别决定	睾丸	否	是	是
RBMY		精子发生	睾丸	是	是	是
AMELY		牙齿发育	牙蕾	否	是	可能
VCY		未知	睾丸	是	是	不适用
PCDHY		未知	大脑	否	是	是

（NRY：Y 染色体上的非重组区域；N/A：不适用）

* 至少 17 个，其中 13 个位于主要假常染色体区域（图 12.15）

尽管在形态上存在区别，X 与 Y 染色体呈现相当区域的同源性，其中包括各类 Xp-Yq 及 Xq-Yp 的同源性，以及 Xp-Yp 及 Xq-Yq 的同源性等。这些同源性导致了各种 X-Y 基因对的发现（图 12.14）。这类同源性的存在提示两条染色体起源于一对原始的同态染色体。很明显，两条染色体在随后发生了相当大的歧化，在一条染色体上位置靠近的序列在另一条上可能相隔很远。X 与 Y 亦能在男性的减数分裂过程中配对，因此能像同源的常染色体那样交换序列。然而，这种减数分裂期的交换在程度上非常有限，仅限于染色体两端很小的**假常染色体区**（pseudoautosomal region，这类序列因此既非 X 连锁亦非 Y 连锁，因而得名假常染色体）。人类有两个假常染色体区。

- **主要假常染色体区**（major pseudoautosomal region，PAR1）在 X 与 Y 染色体短臂最顶端，延伸超过 2.6 Mb，已知含有至少 13 个基因（Ried *et al.*，1998；Gianfrancesco *et al.*，2001）。这是在男性减数分裂中发生必然交叉的部位，被推测为正确的减数分裂分离所必需。这个很小的区域在进化上不稳定（下面），并含有高度重组基因的序列（其性别平均重组率为 28%，对于一个长仅 2.6 Mb 的区域来说，达到了约 10 倍于正常重组率的水平）。诚然，这一高重组率在很大程度上是缘于发生在男性减数分裂中的必然交叉，导致交叉发生率接近 50%。主要假常染色体区与性别特异性区域均被定位于 *XG* 血型基因之内，*SRY* 男性决定基因则位于 Y 染色体上距 PAR1 边界仅 5 kb 的位置（图 12.15）。
- **次要假常染色体区**（minor pseudoautosomal region，PAR2）跨越了 X 与 Y 染色体长臂最顶端约 330 kb 的区域，仅含有四个基因（Ciccodicola *et al.*，2000；Charchar *et al.*，2003）。该区域中 X 与 Y 的交叉并不像 PAR1 那样频繁，而且并非男性减数分裂顺利进行的充分或必要条件。

12.2.7　人类性染色体起源于常染色体，且因周期性区域内的重组抑制而异化

独特的性染色体单独起源于许多根本不同的进化种系的动物中，除哺乳动物外，还包括鸟类（雌性为 ZW，即异配子性别，雄性为 ZZ，即同配子性别），某些种类的鱼、爬行类以及昆虫等。在各物种中，不同的性染色体被推测起源于基本上一样的常染色体，只是它们中的一条偶然演化出一个主要的性别决定基因座来（人类的 *SRY* 基因座，位于 Y 染色体上距 PAR1 边界 5 kb 的位置）。随后的进化导致这两条染色体越来越不相像，在许多物种中，一条性染色体（哺乳动物的 Y，鸟类的 W）缩小成一条短染色体，富含重复序列，而仅含有很少的功能基因。似乎是进化压力导致了具有两条在结构和功能上皆不相同的性染色体的策略。

假常染色体区域的常染色体起源及可塑性

假常染色体区域并未在进化中被很好地保留。在小鼠乃至于某些灵长类中，并不存在 PAR2 的对应序列，已知的小鼠 PAR2 基因的种间同源体或者位于常染色体上，或者被定位于 X 染色体上靠近着丝粒的位置。对 PAR1 来说，亦存在非常显著的物种差异。在人类中，PAR1 的边界位于 *XG* 基因内，而后者在小鼠中似乎并无种间同源体（图 12.15）。对等的小鼠假常染色体区（PAR），长仅 0.7 Mb，并位于 X 染色体长臂的顶端，呈现很低的与人类 PAR 的序列一致性（Perry *et al.*，2001）。小鼠 PAR 的边界位

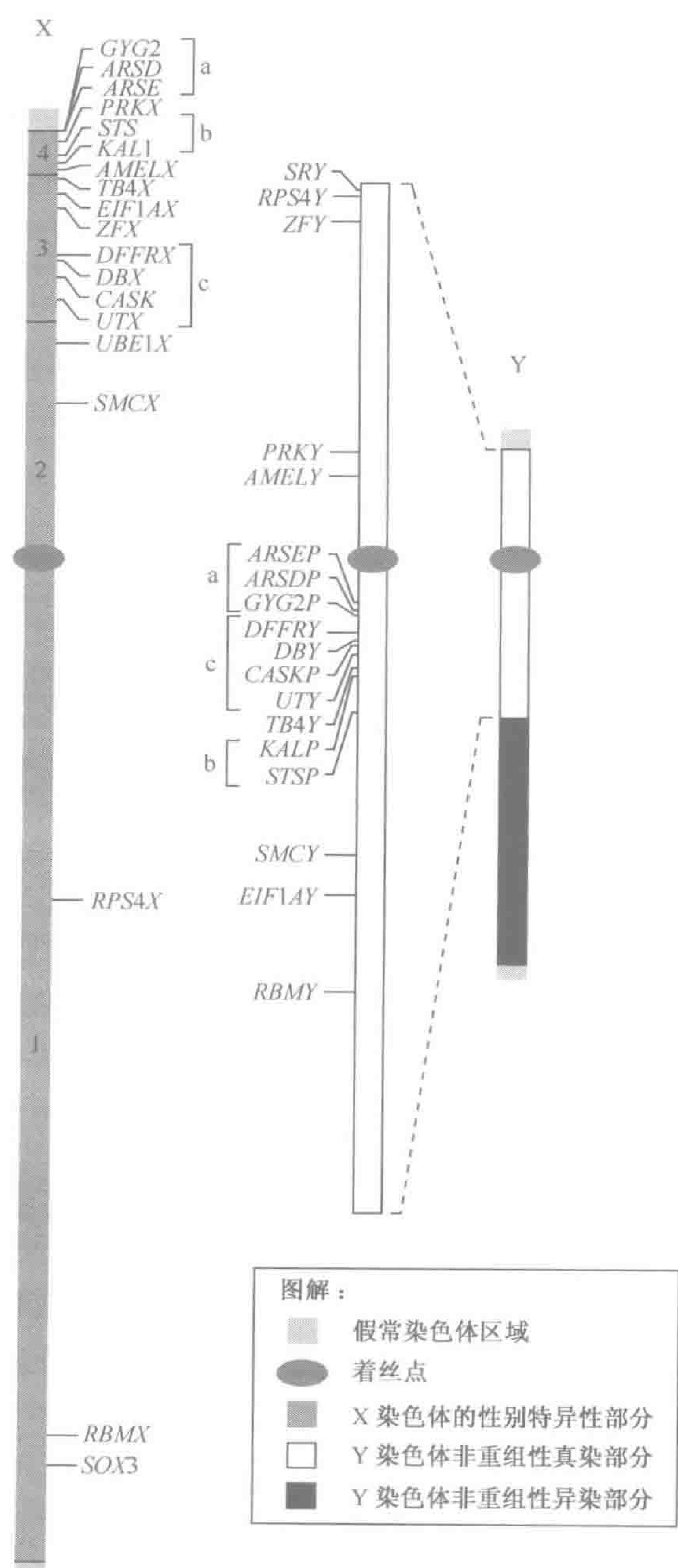


图 12.14　人类的 X 与 Y 染色体呈现若干区域的同源性，提示一个共同的进化起源

与位于 Xq 和 Yq 顶端者一样，位于 Xp 与 Yp 顶端的假常染色体区域完全相同（图 12.15）。其余的非重组区域呈现若干明显的同源性 XY 基因对，加上 *SOX3-SRY* 基因对（二者均具有一个 HMG 结构域）。X 染色体上的数字 1～4 对应于不同的“进化层”（正文）。一些 Y 染色体的同源体已蜕变为假基因（符号以字母 P 结束，如 *ARSEP*、*ARSDP* 等）。经 American Association for the Advancement of Science 允许，复制于 Lahn 和 Page(1999). Science 286，964～967。

于 *Mid1*（旧称 *Fxy*）基因内，其人类种间同源体 *MID1* 则位于更近端的 X 染色体特异性区域。类固醇硫酸酯酶 *Sts* 是唯一已知的位于小鼠 PAR 区域的其他基因，但其人类

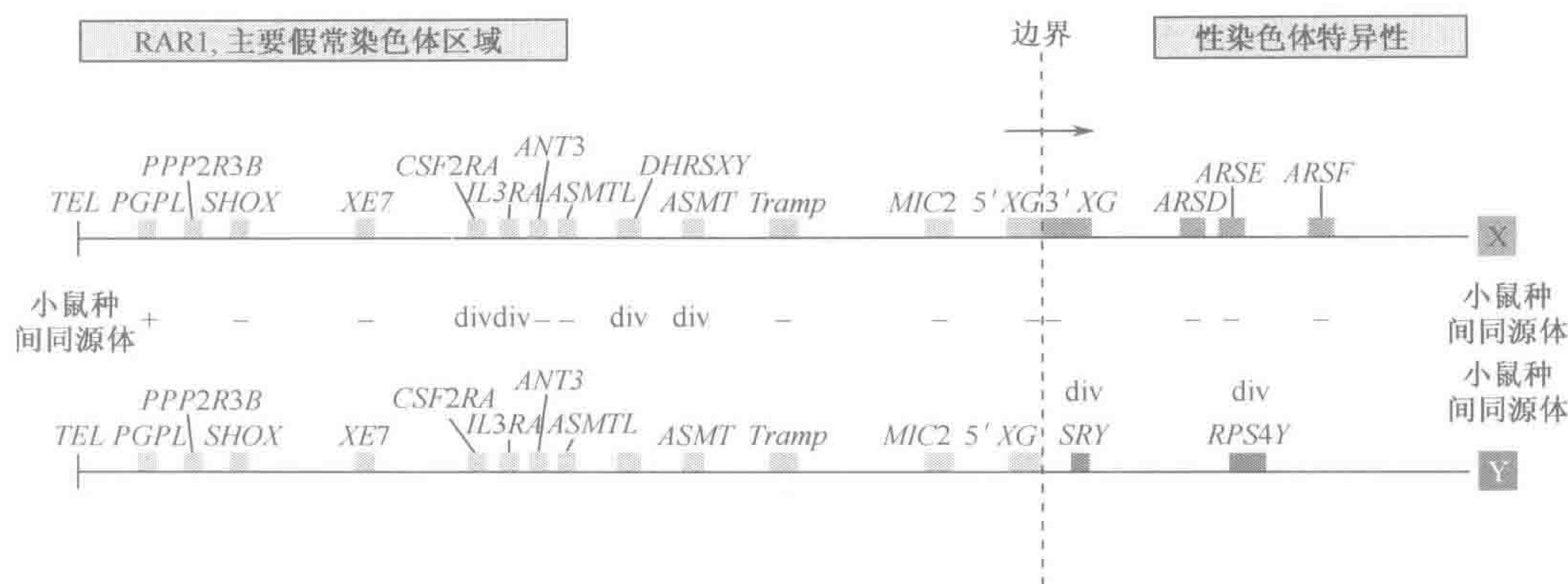


图 12.15　主要的人类假常染色体区（PAR1）的结构与进化学不稳定性

长 2.6 Mb 的 PAR1 区域为 Xp（左上）与 Xp（左下，TEL＝端粒）顶端所共有，且含有至少 13 个基因。紧邻 PAR1 的是 X 与 Y 染色体上较长的性别特异性部分，其中紧邻 PAR1 的基因被标注于边界线右侧。PAR1 的边界位于 X 染色体上的 *XG* 血型基因内。在 Y 染色体上，有一个截短的 *XG* 基因同源体：启动子与前几个外显子存在于假常染色体区域，但随后就是无关的 Y 染色体特异性序列，如 *SRY*、*RPS4Y* 等。PAR1 及邻近的性别特异性区域（早先的一个假常染色体区域的部分）内的基因并未在进化中被很好地保留，小鼠的种间同源体通常难以检测（—）或已高度歧化（div.）。

种间同源体却位于 X 染色体上 PAR1 近端 3.5 Mb 处。人类 PAR1 基因中的三个在某些哺乳动物中具有常染色体种间同源体，其余则在另外一些哺乳动物中具有常染色体种间同源体。PAR1 区域因此被推想为进化于在被重组到另一条性染色体上之前，常染色体片段向两条性染色体之一上的假常染色体区域的反复添加（Graves *et al.*,1998）。

PAR1 与邻近的区域被认为是相对不稳定的区域。频繁的 DNA 交换造成了频繁的基因融合、外显子复制以及外显子混编等（Ried *et al.*,1998）。PAR1 以及邻近的性别特异性区域（原来的假常染色体区——下面）内的许多基因在小鼠中似乎并无明显的种间同源体（根据对已有小鼠基因组序列的分析或杂交反应）。那些具有种间同源体的基因通常正在发生非常迅速的序列歧化，正如位于距 PAR1 边界以及类固醇硫酸酯酶基因 *STS* 仅 5 kb 的主要男性性别决定基因 *SRY* 一样。

人类 Xp 区域的常染色体起源与 X 染色体上的进化层

对亲缘较远的哺乳动物所具有基因的比较提示人类 X 染色体短臂的许多部分来源于最近发生的 X-常染色体易位。哺乳动物被分为两个亚纲，即**原兽**（prototheria，单孔类或卵生哺乳类）与**正兽**（theria），后者又被分为两个亚类，即**后兽亚纲**（metatheria，有袋食肉目）和包括胎生哺乳动物在内的**真哺乳亚纲**（eutheria）。许多真哺乳动物的 X 连锁基因在有袋食肉动物中亦为 X 连锁。然而，定位于一大段人类 Xp（Xp11.3 以远）上的基因却在有袋食肉目以及单孔类中具有常染色体上的种间同源体。由于原兽亚纲的歧化要早于后兽亚纲-真兽亚纲的分离（图 12.24），最简单的解释就是在真兽类种系的早期进化中，至少有一大段常染色体区域被易位到了 X 染色体上。

在共有的 PAR1 和 PAR2 之外，仍存在约 20 对可能属于古老的 X 与 Y 染色体之间

大范围一致序列的遗迹的 X-Y 基因对。在这些 X-Y 基因对中，大部分位于 X 上的基因位于 Xp 上，而其相应序列则似乎见于整个 Y 染色体的常染色质部分（图 12.14）。然而，这些 X-Y 基因对之间亦存在明显的区域差异。以 *Ks*（各同义位点上同义替换次数的平均值，见节 11.2.5）作为序列歧化的指标，X-Y 基因对可被分为四类歧化（年龄）序列，后者与 X 染色体上线性分布的区域相对应（Lahn and Page，1999，表 12.3）。

表 12.3　同源性 X 及 Y 连锁基因可以被分为四类分化序列，对应于 X 连锁基因在 X 染色体上的位置（图 12.14）

X-Y 基因对 （*-假基因）	*Ks*	DNA 差异(%)	X-Y 基因对 （*-假基因）	*Ks*	DNA 差异(%)
第 1 组			第 2 组		
RPS4X/Y	0.97	18	*UBE1X/Y*	0.58	16
RBMX/Y	0.94	29	*SMCX/Y*	0.52	17
SOX3/SRY	1.25	28			
第 3 组			第 4 组		
TB4X/Y	0.29	7	*GYG2/GYG2P**	0.11	7
ELF1AX/Y	0.32	9	*ARSD/ARSDP**	0.09	7
ZFX/Y	0.23	7	*ARSE/ARSEP**	0.05	4
DFFRX/Y	0.33	11	*PRKX/Y*	0.07	5
DBX/Y	0.36	12	*STS/STSP**	0.12	11
*CASK/CASKP**	0.24	15	*KALI/KALP**	0.07	6
UTX/Y	0.26	12	*AMELX/Y*	0.07	7

表 12.3 中的数据表明沿 X 的非重组部分，从长臂末端（Xq 远端）到短臂末端（Xq 远端），X-Y 基因对的年龄呈现逐步下降的模式。明显存在至少四个进化层，对应于图 12.14 中的四组基因，X 与 Y 染色体之间的重组被认为，自第一层即 2.4 亿～3.2 亿年前（产生哺乳类与鸟类的种系发生分离后不久，系统发生信息见图 12.24）开始，曾受到局部范围的抑制。之后，对重组的抑制以独立的步骤扩展到第二层，之后是第三层，最后到第四层。重组被抑制的最可能途径为倒置（据知能够在哺乳动物中对广泛区域内的重组产生抑制），后者似乎是发生于 Y 染色体上（例如，一种 Y 特异性倒置可能解释为何 *PAR1* 的边界跨越了一个完整存在于 X 染色体上，但却破损于 Y 染色体上的基因，见图 12.15）。图 12.16 示意了性染色体进化过程中可能曾经发生的一系列事件。

12.2.8　性染色体的分化导致了 Y 染色体的逐步退化以及 X 染色体的失活

人类 Y 染色体可能正在走向灭绝但男性仍将具有一个未来！

由于重组所提供的遗传新颖性，性别决定系统的进化曾在复杂多细胞生物的发展中扮演关键角色。当一个主要的性别决定基因座在进化过程中形成之后，对含有该基因座的区域中重组的抑制至关重要（以维持性别的差异）。与能够在女性减数分裂期与伴侣 X 染色体在其全长范围发生重组的人类 X 染色体不同，人类的 Y 染色体被视作一个实质上无性的（非重组性）染色体。

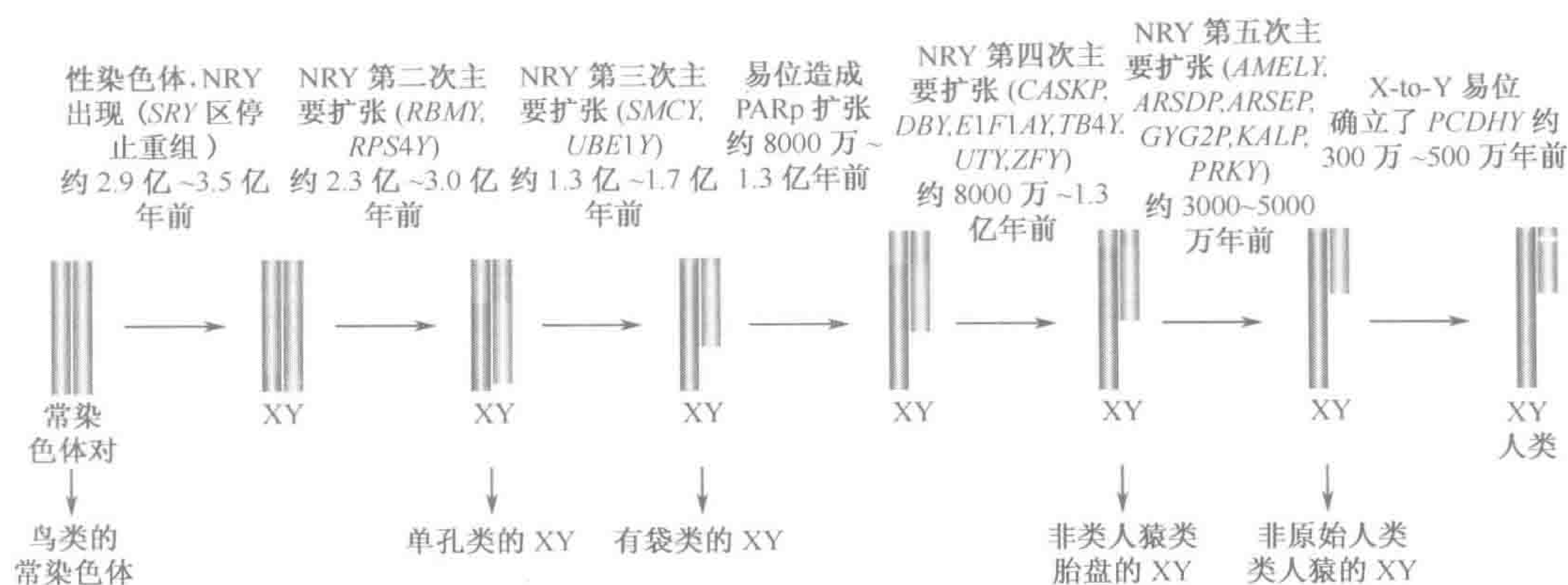


图 12.16 人类性染色体的进化

本图示意可能缘于连续发生的大范围倒置，Y染色体在整体上的缩小及其非重组区域（NRY）的块状扩张。括号中为进化学上新的NRY基因，其系统发生学分支如箭头所示。绿色区域可发生自由重组。红色区域为X染色体特异性。蓝色区域为Y特异性（NRY）。黄色区域表示由X染色体易位至NRY（其他潜在的易位从略）、含有*PCDHX/Y*(protocadherin X/Y）的序列。本图并未按比例绘制，着丝粒被省略，因为其位置在进化的许多阶段中并不清楚。缩写：PAR1，主要假常染色体区域。经Nature Publishing Group允许，复制于Lahn等(2001), Nature Rev. Genet. 2，207～216。

群体遗传学预测一个非重组性染色体将通过一个被称为 **Muller 轴线**（Muller's rachet）的过程逐渐退化。如果突变率足够高，重组的缺乏将意味着在漫长的进化时光里，有害的突变将逐渐在位于该染色体上的基因中积累（因为并无可能通过交换获得一段不含有有害突变的等位序列作为代替）。由于携带较少突变的Y染色体可能被偶然丢失，或者由于它们可能随某个受保护而不发生重组的优势等位基因“搭便车”，突变的等位基因将可能漂移至固定下来。

一旦突变在非重组性Y中蓄积起来并导致基因功能的丧失，保持相关DNA片段的选择压力即不复存在。DNA更新机制将确保该染色体将逐步但无情地被一系列缺失所压缩。其结果就是，人类Y染色体可能正在走向灭绝。然而，男性性别仍将通过转换至另一种性别决定系统而持续存在。最可能的情况就是，它将直接由X∶常染色体的基因剂量比来决定，XO型的个体将成为男性（就像果蝇那样）。

X 染色体失活的必然出现

哺乳动物性别决定机制的进化与提供剂量补偿的X染色体失活机制的形成具有密不可分的联系（节10.5.6；Ellis，1998）。与Y染色体序列的大规模破坏相适应，可能曾出现使X染色体上基因表达增强的压力。但是，这将造成X染色体基因在女性的过度表达，后者则可能导致适合度的下降。其结果就是，一种基因剂量补偿的形式出现了，在女性的细胞里一条X染色体被选择而失活（**X 染色体失活**，X-inactivation）。

X染色体失活的原理就是为那些在Y染色体上没有同源体的X染色体基因提供一种剂量补偿机制。然而，一小部分人类X染色体基因的确具有位于Y染色体上的功能性同源体。由于这些基因并不呈现不同性别之间的剂量差异，它们理应逃逸X染色体失活。所有已测的*PAR1*基因逃逸了X染色体失活。*PAR2*基因则有所不同。两个最

靠近端粒的基因，*IL9R* 与 *CXYorf1* 逃逸了失活，但两个近侧的基因，*SYBL1* 与 *HSPRY3* 则均被失活。这种明显的不一致性是缘于对于这些基因的补偿性 Y **染色体失活**（Y-inactivation）的机制：当出现于 Y 染色体上时，*SYBL1* 与 *HSPRY3* 均被甲基化而不表达。

除假常染色体区域内的基因之外，X 染色体上的所有基因中可能约有五分之一将逃逸失活（Carrel *et al.*,1990）。逃逸的基因主要倾向于在 Xp 上聚集成簇，而且它们中的许多似乎来源于最近发生的常染色体向性染色体的添加。那些在 Y 染色体上没有功能性拷贝、但仍逃逸失活的基因可能属于 2：1 的剂量差异并不引起问题者。非假常染色体区表达模式的物种间差异亦可能存在。例如，人类非假常染色体区基因如 *ZFX*、*RPS4X*、*UBE1* 等均可逃逸失活，但其小鼠同源体如 *Zfx*、*Rps4*、*Ube1X*（并不像人类 *UBE1* 基因那样具有一个 Y 上的同源体）等均发生失活。对于那些剂量需要被严格控制的 X 连锁基因，X 染色体失活的出现似乎是为了适应同源性 Y 连锁基因的蜕变（Jegalian and Page，1998）。

12.3　分子系统发生学与比较基因组学

分子系统发生学是指通过比较核酸或蛋白质来确定有机体或群体之间的进化关系。当然，这将为传统的基于活体生物及收集自化石标本的信息的解剖及形态学特征的系统发生学提供补充。

直到最近，当分子进化生物学家们比较 DNA 序列（或由此推断的蛋白质序列）时，他们通常利用来源于最多不超过几个基因组位置的序列。然而，基于很小序列数据集的系统发生分析可能具有误导性并产生不一致的结果。例如，对于涉及转录、翻译、DNA 复制以及修复的蛋白质的比较突出了藻类与真核生物之间的相似性，而对细胞代谢相关蛋白质的比较则将藻类与细菌分为一组。为了获得更为广泛的观察，他们借助了比较遗传图谱（O'Brien *et al.*,1999）或比较核型分析。各种基因组计划当然地改变了所有这些。它们允许我们从全面的角度来观察整个基因组序列，一个新的学科因此诞生——**比较基因组学**（comparative genomics）。

基因组计划所喷涌出来的序列信息为推断系统发生提供了在权威性及代表性方面要好得多的 DNA 特征。通过确定一整套基因组的序列（第 8 章），我们正在取得对于进化中基因组如何被塑造深刻得多的了解。诚然，这些数据亦可帮助我们了解现存的基因组之间具有怎样的相互关系，并有助于发现重要的保守序列。

12.3.1　分子系统发生学运用序列比对来构建进化树

要构建一棵进化树，就需要比较核酸序列（或者偶尔为推断的蛋白质序列；核酸序列包含更丰富的信息，但在比较亲缘关系很远的基因时，常常使用蛋白质序列）。如果两条或更多的序列呈现足够水平的相似性（**序列同源性**，sequence homology），它们可被假定为来源于一条共同的原始序列。这时，序列比对可被用来获取描述序列间亲缘关系远近的量化评分。

在有足够高的序列同源性的情况下，比较具有相等的固定长度的序列通常较为简

单。然而，被比较的序列常常已发生缺失或插入，因此需要采用严格的数学方法来进行序列比对。各种辅助性的算法已经被设计出来。Needleman 与 Wunsch（1970）的**相似法**（similarity approach）寻求将匹配核苷酸的数目最大化；Waterman 等（1976）的**距离法**（distance approach）的目标则是使错配的数目最小化。诸如 Clustal 之类的计算机程序能够同时对多条序列进行比对（Jeanmougin *et al.*,1998）。

序列一旦被比对，**进化树**（evolutionary tree）即可被构建。它们最常表示为由线条（**分支**，branch）与节点所构成的图。被比较的不同生物（序列）位于外侧的节点上，但通过线条与**内部的节点**（interior node，分支之间的交叉点）相连，后者则代表两个或更多生物的原始形态。一棵**有根树**［rooted tree，或**进化分支树**（cladogram)］推测某个共同祖先（表示为树干或树根）的存在，并标明进化的方向（图 12.17)。进化树的根可能通过将序列与**外群**（outgroup）序列（在进化学上具有明确的关联，但仅与被研究的序列具有较远的亲缘关系）相比较而确定。一棵**无根树**（unrooted tree）则不支持共同祖先的存在，仅表示生物之间的进化学关系。值得注意的是，潜在的有根树的数量要远远超过无根树的数量。

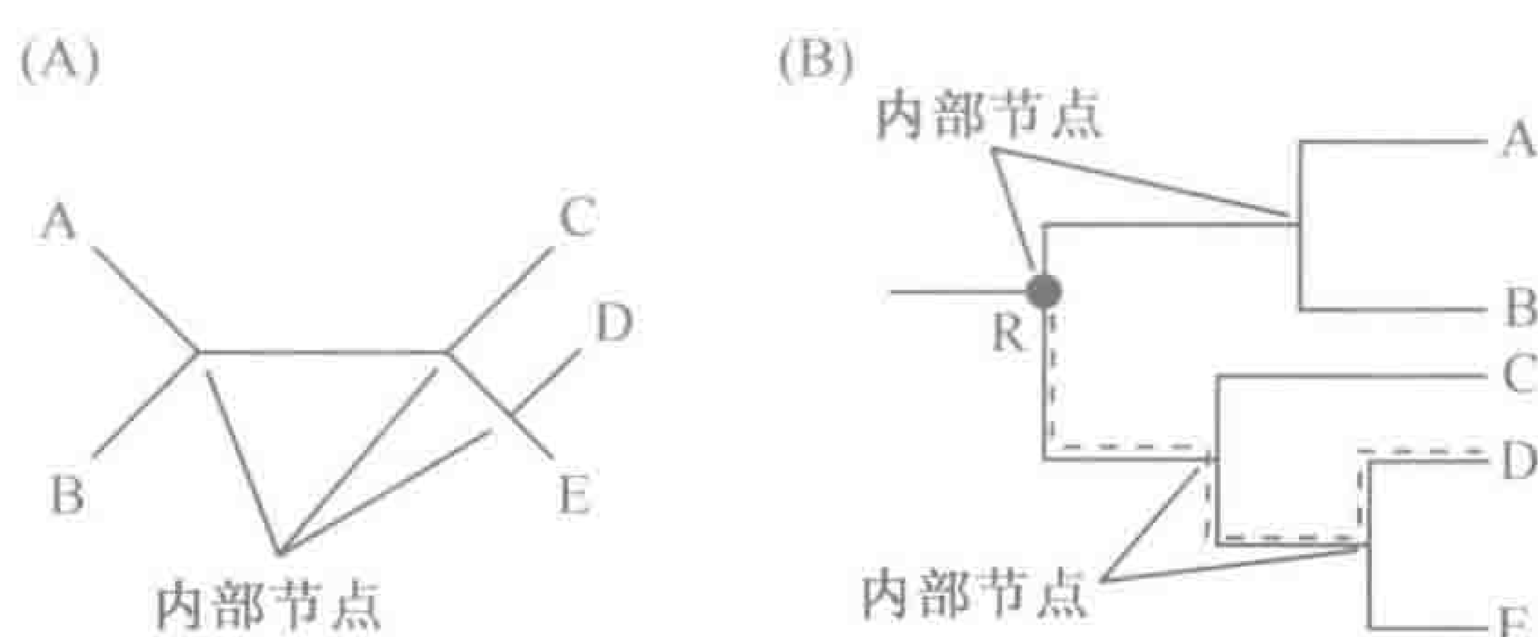


图 12.17　无根与有根的进化树

(A) **无根树**。该树具有 5 个外部节点（A、B、C、D、E)，之间由交叉于内部节点的线条（分支）相连。这样一棵树仅能说明被研究物种之间的关系，但并不能够明确其进化途径。(B) **有根树**。从某个特定的内部节点，发出其根部（由黑色的 R 表示)，通往其他任何节点皆有一条独特的进化途径，例如通往 D 的途径（虚线)。

构建进化树

各种不同的方法被用来构建进化树，但许多人运用一种**距离矩阵法**（distance matrix)。其第一步是计算数据集中所有序列对之间的**进化学距离**（evolutionary distance）并将其排列成一个表（阵列)。这可以被表示为两条序列之间核苷酸差异或氨基酸替换的数目，或者每个核苷酸（氨基酸）位置上核苷酸（氨基酸）差异的数目（以图 12.18A 为例)。

在列出序列对间差异的矩阵之后，下一步就是根据序列之间的进化学距离对其进行联系。例如，在一种方法中，具有最小距离分值的序列对的两个成员被连接起来，并从二者之间引出一条根来。该序列对的成员与第三个节点之间距离的平均值将被用于距离矩阵的下一步计算，这个过程将被重复直至所有的序列均被放入树中。这将总是产生一棵有根树，而不同的方法如**邻近关系**（neighbor relation）法则可能产生无根树。关键的假设在于一个恒定的**分子钟**（molecular clock，在不同种系中突变率恒定不变)，而这常常可能并不正确（节 11.2.6)。**邻接法**（neighbor-joining method）是一种并不要

求所有的种系在相同的单位时间内发生等量分化的方法。因此，它特别适用于由进化速度存在很大差异的种系所构成的数据集（由该方法构建的一棵进化树的例子见图 12.31）。

矩阵距离法的替代方法包括最大节俭法与最大似然法。**最大节俭法**（maximum parsimony）试图使用最少的进化步骤。它们将考虑所有能够解释序列间关系的进化树，之后选择那些需要最少改变者。**最大似然法**（maximum likelihood）则创建所有可能的进化树，之后运用统计学来评估哪棵树最有可能。对于少数序列来说，这将可行，但对于一大堆序列来说，产生出来的进化树数目将变得如此巨大，以至于得运用启发式算法（在一段可计算时间内产生出一个答案的方法，但该答案可能并非最优）来选择进化树的一个子集来构建。

评价一棵进化树的准确性

一旦得到了一棵进化树，即可运用统计学方法对它的可靠性进行评估。一种流行的方法是**自举法**（bootstrapping），即一种 Monte-Carlo 仿真。通常的情况是，数据中的一个子样品被移除，并被替换成一组随机产生的相应数据，由此产生的假序列将被分析，以检查暗示的进化模式是否依然被支持（图 12.18B）。重新采样的过程将数据随机化，但如果在两条序列之间存在明显的联系，随机过程将无法抹杀它。另一方面，如果在最初的进化树中联系两条序列的节点是虚假的，它可能在随机化时消失，因为随机化过程将改变个别位置上的频率。自举法通常涉及重复采样部分数据 1000 次。最高的**自举值**（bootstrapping value）通常被表示为一个百分数，因此数值 100 将意味着仿真结果完全支持最初的判断。数值 95～100 提示所预测的节点具有高水平的可信度。低于 95 的自举值并不意味着对序列的原始分组是错误的，而是现有的数据并不能提供令人信服的支持。一些例子见图 12.31。

12.3.2　新的计算机程序可比对大尺度以及整个基因组的序列，有助于进化分析与保守序列的发现

诚然，从基因组计划流出的序列数据仅仅是找出其涵义这一长期任务的开端。由于这些数据集的巨大尺寸，需要新的计算机程序来进行大规模及整个基因组序列的比对。这些新程序对于发现保守很好的序列以及了解其进化关系具有宝贵的价值。

用于发现保守序列的大尺度序列比较分析

对于基因组计划的一个主要挑战就是找出 RNA 基因、调控序列以及其他重要功能性序列等并不编码蛋白质的序列。对于复杂基因组（低比例的编码 DNA）序列的分析并不容易，因为计算机程序并不能轻易预测 RNA 基因以及调控序列（编码多肽的基因较为容易——因为存在较长的可读框，以及大量有关基因及其表达产物的序列数据提供对比）。一种替代方案就是通过对跨度较大的基因组序列进行比较，以寻找高度保守的序列。不同的计算机程序，如 VISTA（http://www-gsd.lbl.gov/vista）以及 Pipmaker（http://bio.cse.psu.edu/pipmaker/）等均能进行大规模的序列比对，并能以图像格式显示结果（一个例子见图 12.19）。

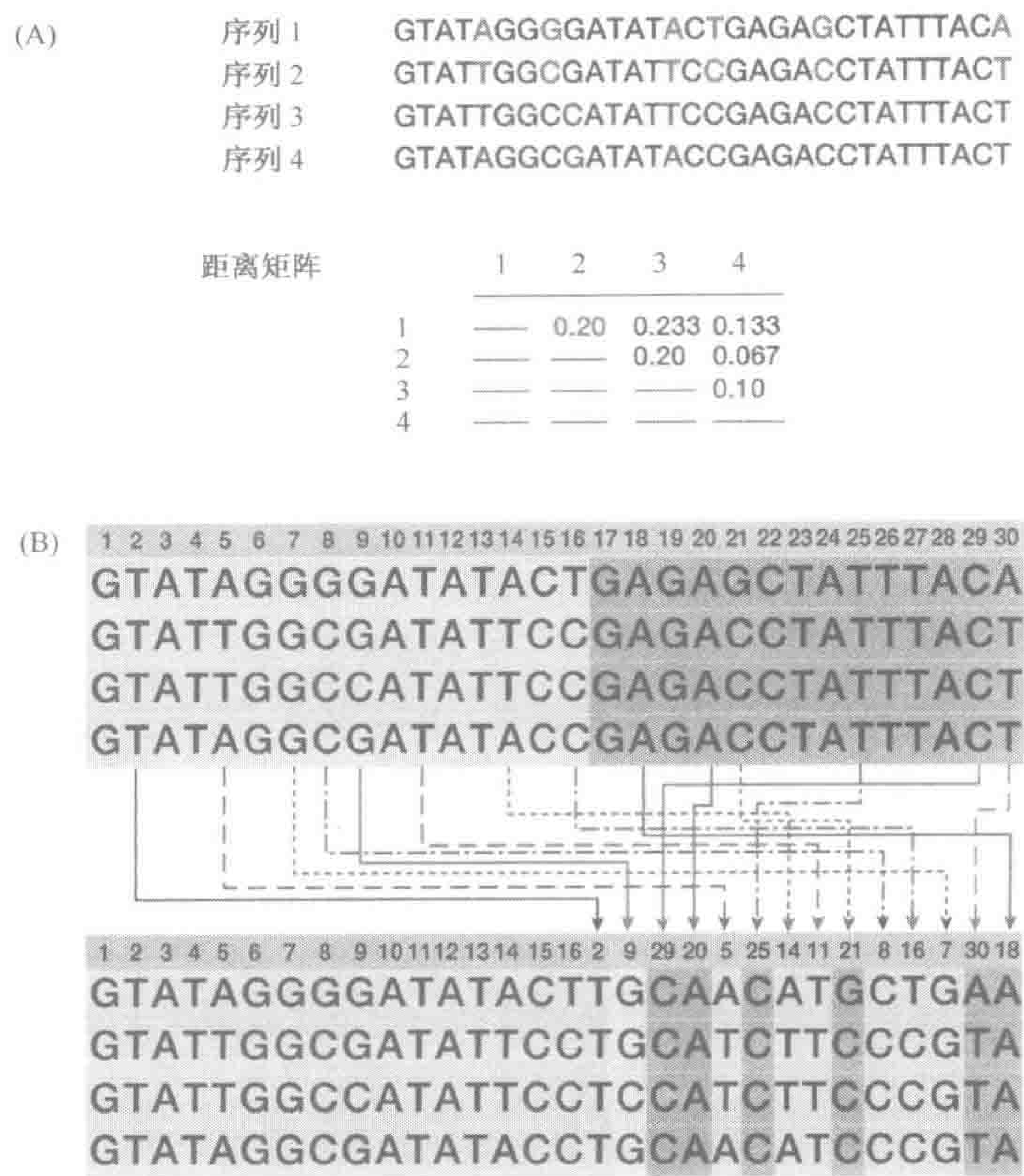


图 12.18　运用距离矩阵法构建一棵进化树，并运用自举法对其进行验证

(A) **构建进化树**。多重序列比对被用来计算序列对之间的进化学距离。序列 1 和 2 在 6/30(= 0.20) 的核苷酸位置上存在差异；3 和 4 在 3/30(= 0.10) 的核苷酸位置上存在差异。这些与其他序列对间的差异数值被录入一个距离矩阵，计算机程序将利用这些数据来估算序列之间的关系并构建一棵进化树。距离矩阵法需要包括对可能发生的多重替换（一个简单的 A→T 替换实际上可能缘于连续发生的 A→C→T 替换）的数量所进行的统计学估计。(B) **自举分析**。原始数据集的一部分（在这里为位置 17～30）被选择以去除，其余（位置 1～16）则被保留。原始位置 17～30 被一个同样大小、从 30 个原始位置上随机挑选位置所替换，产生一个新的假序列比对，其中某些原始位点被再现（2、9、5、14、11 等），某些则消失（17、21、23 等）。这一过程将被重复 1000 次左右，以产生 1000 个假序列比对结果且每次都构建进化树，并与原始结果核对，得到一个自举值（见正文）。

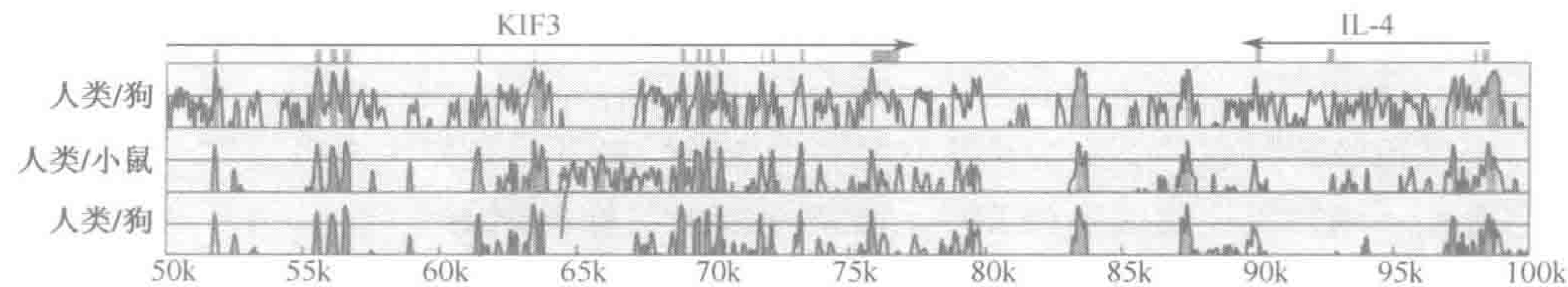


图 12.19　对一段长 50 kb，含有人类 *KIF3* 和 *IL4* 基因的序列与狗（D）以及小鼠（M）的种间同源体所进行的 VISTA 序列比对

保守序列由相对于其在人类基因组中的位置（横轴）以及纵轴所显示的它们之间的一致性百分数（50%～100%）来表示。编码外显子（上方的蓝色方框）以及 *KIF3* 基因的 3′-UTR（青绿色方框）的位置被标注于图的上方。水平箭头表示各基因转录的方向。由高度保守序列所形成的峰包括编码序列（蓝色）以及功能未知的非编码序列（红色）。经冷泉港实验室出版社允许，源自 Dubchak 等（2000）. Genome Res. 10，1304～1306.

作为比较性哺乳动物分析（多数为人-小鼠）的结果之一，人类基因组中高度保守序列的数量目前认为约占 5%，其中仅大约 30%（整个基因组的～1.5%）被认为是多肽编码 DNA。许多分析比较了人类与小鼠的序列，但一些涉及灵长类特异性基因表达的调控序列将可能需要对一系列的灵长类序列进行比较。

模式生物的基因组亦被用于与同属、但亲缘关系较远的物种的基因组序列相比较。例如，线虫属中的秀丽新小杆线虫与 *C. briggsac* 于 1 亿年前分化自一个共同祖先，二者的基因组序列目前均可通过网址为http://www.ensembl.org 的 ENSEMBL 查询。拟暗果蝇（*Drosophila pseudoobscura*）的基因组计划（http://www.hgsc.bcm.tmc.edu/projects/drosophila/）将使其与果蝇基因组序列的比较成为可能。

整个哺乳动物基因组序列的比对

复杂基因组中含有较高比例的中性演化序列。由于缺乏选择压力，这类序列分化迅速。为了获取对于基因组进化的更多的了解，计算机程序需要能对被比较基因组中至少较高比例的中性演化序列进行比对。对于人类及小鼠基因组来说，这已通过对 BLASTZ 程序的改进而实现（Schwartz *et al.*, 2003），**人-小鼠全基因组序列比对**（whole human-mouse genome sequence alignments）以及有用性的示意可通过http://bio.cse.psu.edu/genome/hummus/来查询。

全基因组的比对显示约 40%的人类基因组序列与小鼠匹配，高度重复的 DNA 序列可被分成两类。**种系特异性重复**（lineage-specific repeat）来源于发生在人类与小鼠自共同祖先分化（大致在过去约 8000 万至一亿年之间）之后的转座。来自共同祖先的**古老保守性重复**（ancient conserved repeat）得到了保存，即使它们已有超过 8000 万至一亿年的历史，种间同源性的人类与小鼠重复亦能够被识别并比对。数据亦显示两类重复在比例上的差异。在人类中，24.4%的基因组（或者大约 53%的散在重复）为种系特异性，而在小鼠中，它们约占整个基因组的 32.4%，或散在重复（约占小鼠基因组的 39%）的 85%。小鼠基因组中来自于祖先的重复已在很大程度上被替换，仅占基因组的 5%，而在人类中，22%的基因组由来自于祖先的重复所构成。这些差别反映了两个基因组中核苷酸替换速度的差异（节 11.2.6）。

12.3.3　基因数目通常与生物学复杂性成正比

基因组测序已经揭示出一些令人惊奇的事实，乍一看去，基因数目可能并不与生物学复杂性或者预期相符。有谁会想到我们的基因的数量仅是长 1mm、具有仅约 1000 个细胞的线虫的 1.5 倍，抑或这种小虫子似乎要比复杂得多的果蝇多几千个基因？尽管如此，如果我们观察已经测序的基因组，就会发现一个明显的趋势：脊椎动物的基因组具有近两倍于无脊椎动物的基因，而后者又具有比单细胞生物多得多的基因（表 12.4）。

后生动物的基因数目通常随其特化而增加，但是，在基因复制对复杂的后生动物的出现起到重大推进作用的同时，某些物种，例如果蝇等却具有少得惊人的复制性基因。必须考虑到的是，在进化中基因亦可能从种系中丢失。

由表 12.4 可知，从简单基因组到复杂的后生动物基因组，基因密度逐渐下降。这反映了早期真核生物中内含子的出现，以及随着基因组变得更加复杂，内含子以及基因

之间的区域中重复 DNA 不断增加的蓄积趋势。因此，例如，约 45%的人类基因组由基于转座子的重复所构成，但是在小鼠中，对应的值则是 37%，而果蝇与秀丽新小杆线虫中的相应值则更低。

表 12.4 简单与复杂基因组中的基因数目及密度

单细胞基因组				多细胞基因组			
基因组	基因组尺寸[范围]	基因数	基因密度	基因组	基因组尺寸（或已测序的 Mb 数）	基因数	基因密度
原核生物				**无脊椎动物**			
细菌（n=97）	3.10 Mb [0.58～9.11 Mb]	平均 = 2840	每 1.09 kb 一个	海鞘类（玻璃海鞘，*C. intestinalis*）	117 Mb*	16 000	～每 7 kb 一个
藻类（n=16）	2.23Mb [1.66～5.75 Mb]	平均 = 2200	每 1.02 kb 一个	线虫（秀丽新小杆线虫，*C. elegans*）	97 Mb	19 000	～每 5 kb 一个
单细胞真核生物				果蝇（黑腹果蝇，*D. melanogaster*）	123Mb*	14 000	～每 9 kb 一个
小孢子虫							
脑胞内原虫（*E. cuniculi*）	2.9 Mb**	2000**	每 1.45 kb 一个	蚊子（冈比亚按蚊，*C. gambiae*）	278 Mb	14 000	～每 20 kb 一个
单细胞真菌				**植物**			
酿酒酵母（*S. cerevisiae*）	14 Mb	6300	每 2.2 kb 一个	阿拉伯芥（拟南芥，*A. thaliana*）	115 Mb*	25 500	～每 4.5 kb 一个
裂殖酵母（*S. pombe*）	14 Mb	4800	每 2.9 kb 一个	**脊椎动物**			
				河豚（红鳍东方豚，*T. rubripes*）	365 Mb	31 000	～每 12 kb 一个
原生动物							
恶性疟原虫（*P. falciparum*）	23 Mb	5300	每 4.3 kb 一个	小鼠（小家鼠，*M. musculus*）	2500 Mb*	～30 000?	～每 80 kb 一个
约氏疟原虫（*P. Y. yoeli*）	23 Mb	5900	每 3.9 kb 一个	人	2900 Mb*	～30 000	～每 100 kb 一个

*——并不代表基因组的全长，因为已除去难以克隆及测序的高度重复性串联重复，仅是全部的基因数目

**——较小的基因组尺度以及较少的基因数目反映了小孢子虫必然处于细胞之间的状态

12.3.4 逐渐的蛋白质特化的程度正在为蛋白质组比较所揭示

当已被详尽研究过的模式生物如大肠杆菌和酿酒酵母等的基因组序列被首次确定时，人们惊奇地发现其中有相当多的基因功能尚属未知。对于人类基因组序列草图来说，接近一半具有未知的功能，这提醒我们在真正开始了解我们的基因组之前，还有很长的路要走。在其余者中，接近一半涉及信号传导或者核酸的结合（图 12.20）。

将预测于人类基因组序列的蛋白质与预测于各种模式生物基因组测序结果者进行比较为了解进化过程中基因的使用情况提供了一些线索（图 12.21 以及表 12.5）。大约 20%的人类蛋白质与广泛分布于真核生物及原核生物中的蛋白质呈现序列同源性。约 1%的蛋白质在已分析的动物基因组中并无同源体。然而，当时有关非人灵长类的数据库非常之少，因此上述的 1%左右将可能被证实为灵长类特异性而非人类特异性（至

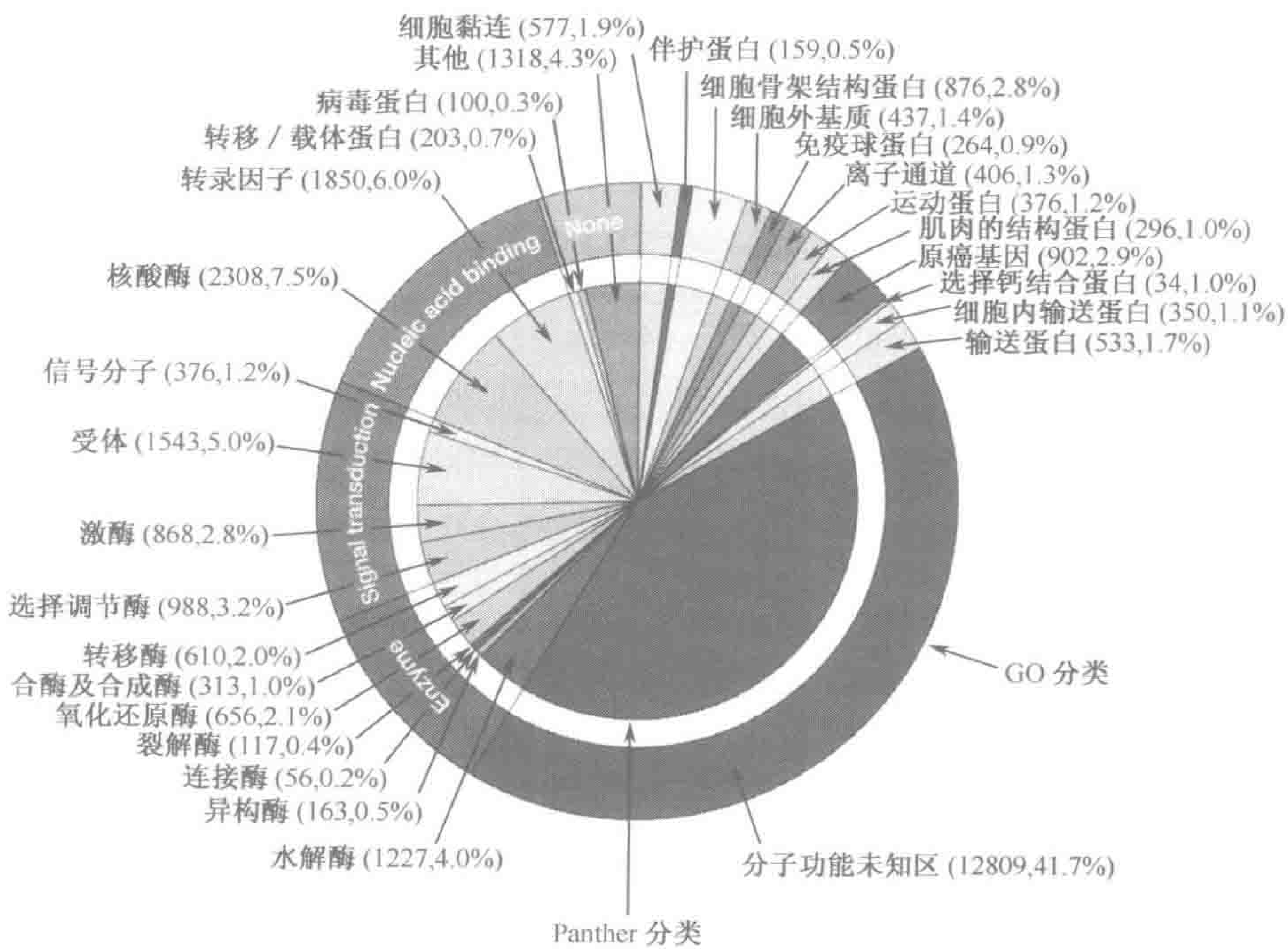


图 12.20　人类多肽编码基因的初步功能分类

26 383 条人类多肽编码基因的已知或预测功能。分类的依据为 GO 分子功能分类，由外侧圆环（基因实体分类，节 8.3.6）或 Celera's Panther 分子功能分类（内侧圆环）。经美国科学促进会允许，复制于 Venter 等 (2001). Science 291，1304～1351。

多，仅有很少的人类基因被推测为人类特异性）。

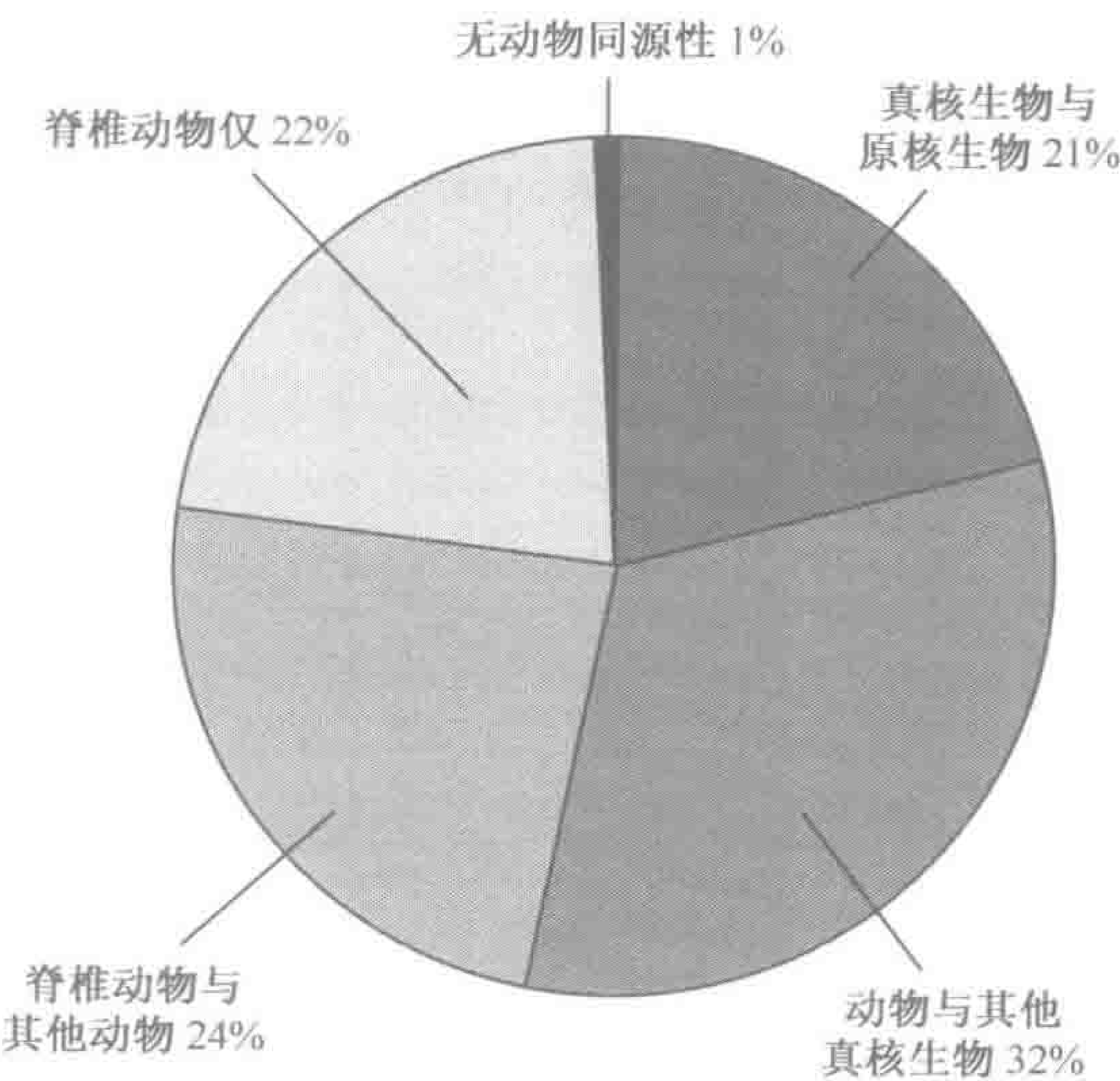


图 12.21　预测的人类与小鼠蛋白的分类学分布

我们的许多蛋白质都具有古老的进化起源，仅不足四分之一起源于脊椎动物出现之后。值得注意的是，新的序列仍在不断被添加到数据库中，真正为人类所专有的蛋白质据推测非常之少。经 Nature Publishing Group 允许，修改自人类基因组测序协作组（2001）。

不同物种的完整基因组的可用性亦允许一种形式的**比较蛋白质组学**（comparative proteomics），即对不同基因组所编码的所有已知或预测的蛋白质进行详细分析，以发现其中的特异性蛋白结构域、基序或其他亚序列结构。领先的资源是由欧洲生物信息学研究所所提供的各种蛋白质序列数据库组合而成的 InterPro（蛋白质家族、结构域以及位点的整合资源）（http://www.ebi.ac.uk/interpro/）。InterPro 所提供的数据具有不同的形式，其中包括对特殊蛋白质家族、蛋白质结构域以及已测序基因组相关蛋白质组中的蛋白质重复等的分类列表（见http://www.ebi.ac.uk/proteome/）。

将前 25 条人类 InterPro 记录（基于普遍性）与不同类型的真核生物基因组进行横向比较提示，若干种类不仅未见于单细胞的酵母（11/25）中，而且未见于植物拟南芥（6/25）中，在无脊椎动物中则程度较轻（秀丽新小杆线虫 3/25，果蝇 2/25，表 12.5）。这在部分程度上是由于脊椎动物的特化，例如免疫系统的基因。在小鼠和人类中，嗅觉受体基因显然很重要，尽管在其他生物中未见与之直接对应的基因，秀丽新小杆线虫的第一个 InterPro 记录即是可能具有嗅觉受体功能的线虫 7TM 化学受体。InterPro 记录在频度排行上亦显示重大的物种差异，即使对于人类和小鼠来说，例如对于嗅觉受体、C2H2 型锌指以及 KRAB 盒（见于 C2H2 型锌指蛋白内的结构域）以及视紫质型 G 蛋白偶联受体等基因的普遍性来说（表 12.5）。

表 12.5　一些蛋白质家族、结构域和重复在后生动物中的分布比较

蛋白质家族结构域或重复	人		小鼠		果蝇		秀丽新小杆线虫		拟南芥		酿酒酵母	
	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行
锌指,C2H2 型	946	1	482	6	360	1	246	11	191	20	54	10
免疫球蛋白/主要组织相容性复合体	883	2	674	3	160	8	120	26	51	110	None	N/A
类视紫质 GPCR 超家族	826	3	1375	1	86	26	403	4	6	836	None	N/A
类免疫球蛋白	804	4	655	4	144	10	100	35	1	1920	None	N/A
蛋白激酶	687	5	486	5	263	2	507	2	1047	1	118	1
嗅觉受体	477	6	980	2	None	N/A	None	N/A	None	N/A	None	N/A
丝氨酸/苏氨酸蛋白激酶	472	7	319	7	205	4	286	7	816	2	113	2
G 蛋白 βWD-40 重复	386	8	243	10	188	5	166	17	267	12	101	3
锌指,RING	367	9	222	13	121	15	173	16	466	5	39	16
类 EGF 结构域	357	10	286	8	101	21	202	13	34	175	None	N/A
RNA 结合区域 RNP-1(RNA 识别结构域)	342	11	234	12	166	7	160	18	299	10	58	8
类 Pleckstrin	331	12	188	18	79	35	90	46	31	197	29	25
免疫球蛋白亚型	327	13	191	16	79	35	28	161	None	N/A	None	N/A
富含脯氨酸的伸展蛋白	324	14	185	19	147	9	150	19	186	21		N/A
酪氨酸蛋白激酶	314	15	216	15	132	12	192	14	360	7	26	28
KRAB 盒	314	15	119	32	None	N/A	None	N/A	None	N/A	None	N/A
钙结合 EF-手	298	17	22	14	128	13	134	23	220	16	17	46
锌指,C2H2 亚型	286	18	91	48	2	1116	2	1161	None	N/A	None	N/A
SH2 结构域	280	19	189	17	80	33	70	64	4	1075	25	29

续表

蛋白质家族结构域或重复	人		小鼠		果蝇		秀丽新小杆线虫		拟南芥		酿酒酵母	
	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行	匹配蛋白	排行
锚蛋白	259	20	162	20	94	23	110	27	114	38	19	42
免疫球蛋白 C-2 型	259	20	145	26	111	17	67	67	None	N/A	None	N/A
同源基因盒	254	22	238	11	107	19	106	30	95	53	8	125
纤连蛋白，Ⅲ型	232	23	157	22	71	42	54	78	4	1075	2	584
免疫球蛋白Ⅴ型	223	24	249	9	4	713		N/A	None	N/A	None	N/A
富含亮氨酸的重复	212	25	146	25	108	18	69	65	509	4	6	183

物种名称下方的各种蛋白质数目源自 SWISS-PROT、TrEMBL 以及 Ensembl 所收录的非冗余性蛋白质组。左边的纵栏列举了 InterPro 数据库中前 25 种人类条目（根据匹配蛋白的数量）。物种名称下方为各种匹配蛋白的数量及排行。数据来自欧洲生物信息学研究所（2003 年 3 月），网址为http://www.ebi.ac.uk/proteome。

12.4　我们因何而变成人？

我们从哪里来？我们因何而变成人？诸如此类的根本性问题在我们拥有了现有生物的完整基因组、基因名单以及基因产物的详细目录之后将会得到最好的答案。自 1995 年起，基因组计划即开始交付基因组序列，在新千年的头 10～20 年中，将会看到被测序的基因组、详细的基因注解以及有关蛋白质组及 RNA 产物的全面信息的爆炸式增长。对于可用数据的比较分析已经开始为我们提供诸如我们与其他物种共同具有哪些核酸序列和预测蛋白，以及我们如何与之不同之类的分子水平的丰富信息。

对比当然可以在不同的水平进行，人类可以被放在各种系统发生群中。在一系列进化上不断特化的种群中，我们同时属于真核生物、后生动物、左右对称生物、有体腔生物、后口动物、脊索动物、有头盖动物、脊椎动物、有下颌动物、羊膜动物、真哺乳纲动物、狭鼻猿类、类人猿、原始人类等（图 12.22～12.24，以及框 12.5 中的词汇表）。

遗传密码的统一性，关键生化反应以及细胞功能中核心步骤的极大的进化保守，以及后生生物细胞中某些关键发育过程的高度保守等，均为强调人类与形态迥异、且进化关系遥远的生物之间具有密切关系的特征。当然，亦存在许多差别。更为复杂的后生动物倾向于具有含有更多基因及蛋白质结构域以及更多重复序列的复杂基因组。不同的物种在基因表达的有关特征方面亦存在显著的差异。具体的例子包括：

- **操纵子**——与细菌操纵子的协同调控不同，真核生物基因的转录倾向于具有独立调控，然而秀丽新小杆线虫的一大部分基因具有类似于细菌操纵子的结构（但在具体操纵上有所不同）；
- DNA **甲基化**（在后生动物中远非一致，框 9.3）；
- **基因组印记**（哺乳动物基因表达的一个重要特征，但似乎不存在于诸如果蝇、线虫之类的生物中）；
- **脊椎动物的适应性**，如免疫相关，激素等；
- **哺乳动物的适应性**，如 XY 性别决定机制相关基因，X 染色体失活，胎盘形成等。

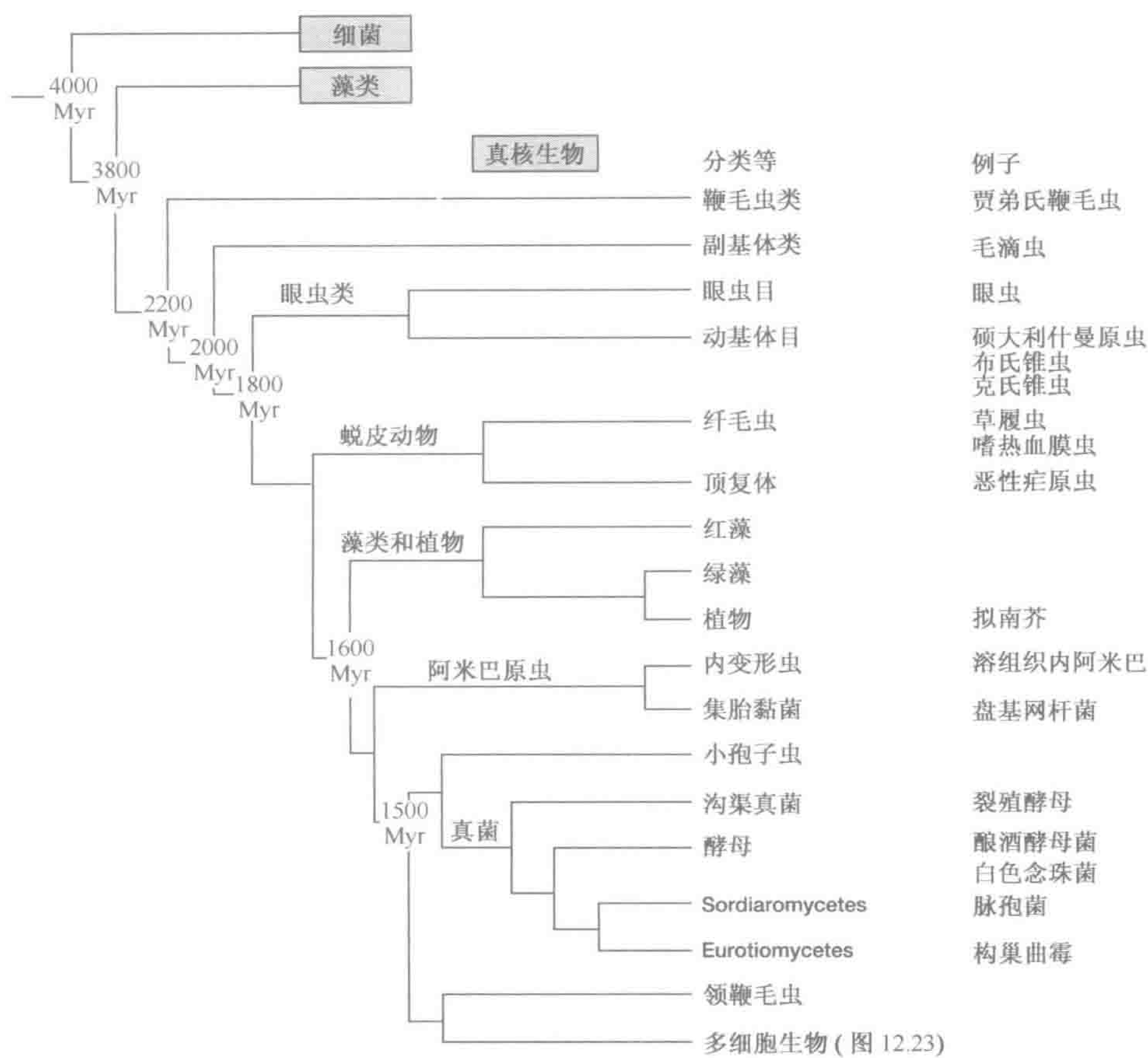


图 12.22 一个简化的真核生物系统发生图

注：领鞭毛虫（choanoflagellate）被认为是与海绵亲缘最近的现存原生生物和最原始的后生动物（领鞭毛虫在性状和功能上几乎与海绵的环细胞，或称襟细胞完全相同）。

诚然，我们亦不同于其他哺乳动物以及我们在进化上的表亲如大型的猿类等。最近对于小鼠与大鼠基因组的测序以及正在进行的黑猩猩（Olson and Varki，2003）及其他灵长类基因组测序等，将为了解它们的结构及其与人类之间的差异提供重要信息。

12.4.1 是什么使我们区别于小鼠？

最近对于小鼠基因组测序的一份评论提议人类最好的朋友已不再是狗，而是小鼠。毋庸置疑，小鼠是最为重要的哺乳动物模型（其优点参见框 8.8），在某些领域里，我们严重依赖于通过小鼠研究进行推断（例如研究基因在发育中的表达）。诚然，小鼠仍仅仅是一个模型，我们也越来越意识到人类与小鼠之间的差别。以下的观察资料中的大部分均来自小鼠基因组测序协作组（Mouse Genome Sequencing Consortium）（2002），在下面的对照之中简称为 MGSC(2002)。

基因组结构的概况

小鼠基因组的常染色质部分大小为 2500 Mb，比人类基因组长约 2900 Mb 的常染

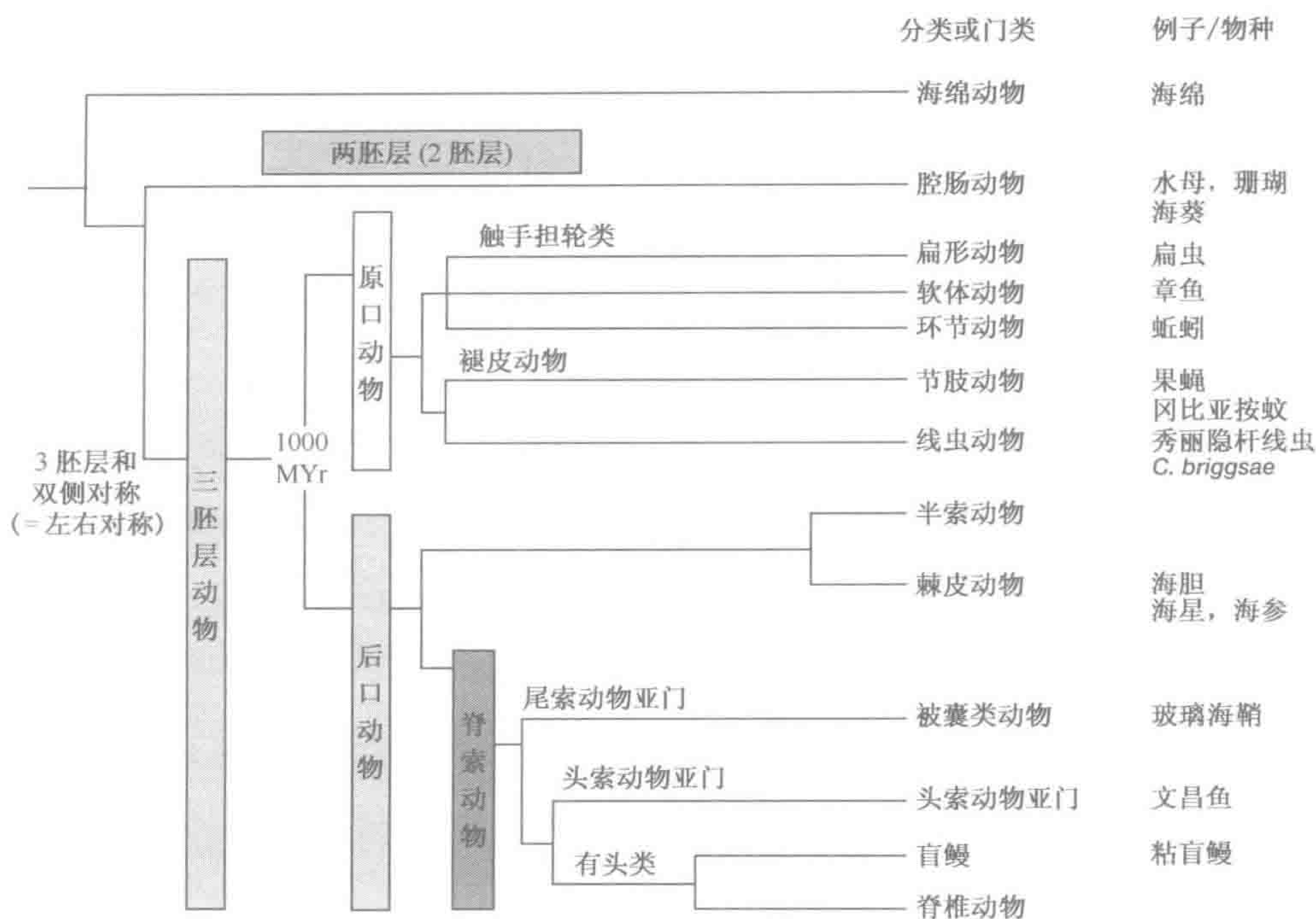


图 12.23　一个简化的后生动物系统发生图

注：原口动物-后口动物发生于大约 10 亿年前的原始性分离意味着秀丽新小杆线虫和果蝇与人类之间的亲缘关系较其他无脊椎动物如海胆（*Strongylocentrotus purpuratus*），被囊动物如玻璃海鞘（*Ciona intestinalis*），以及头索动物如文昌鱼（*Amphioxus*）等更远。

色质部分略短。这一差异在很大程度上是缘于后者中数量更多的重复性序列（见下文）。因此，小鼠基因的内含子平均约短 16%。基因之间的距离一般更短，但在区域之间存在相当大的差异。外显子大小和编码 DNA 的大小（平均 500～550 个密码子）在人类与小鼠中非常相似，种间同源基因所含的外显子数目亦然。小鼠基因组中的总（G+C）含量为 42%，略高于人类基因组 41%的总（G+C）含量。不过，人类基因组具有显著较多的片段，它们的 DNA（G+C）含量很高（MGSC，2002，图 7），因此其中的 CpG 岛（在不含重复的人类 DNA 中共有～27 000 个）亦远多于小鼠（15 000 个）。保守的人-小鼠同线性区域平均长约 10 Mb（图 12.11）。

散在的重复

人-小鼠基因组大小的差异在很大程度上是缘于人类基因组中较多的重复序列：基于转座子的重复占人类基因组的～45%，但仅占小鼠基因组的～37%。人类 LINE 序列的数量略多于小鼠的 LINE 序列。虽然 LINE 序列继承于人类与小鼠的共同祖先，小鼠基因组却具有更高的 LINE 序列更新率。在两种基因组中，仅 LINE1 元件仍在活跃地转座，但现代小鼠 LINE1 序列中的绝大多数在人-小鼠分离之后发生了转座，而大多数人类 LINE1 重复序列则从共同祖先那里保存了下来。LINE-1 重复在人类与小鼠（以及其他所有哺乳动物）中得到了保留，主要是由于保守性选择压力以维持编码反转录酶的

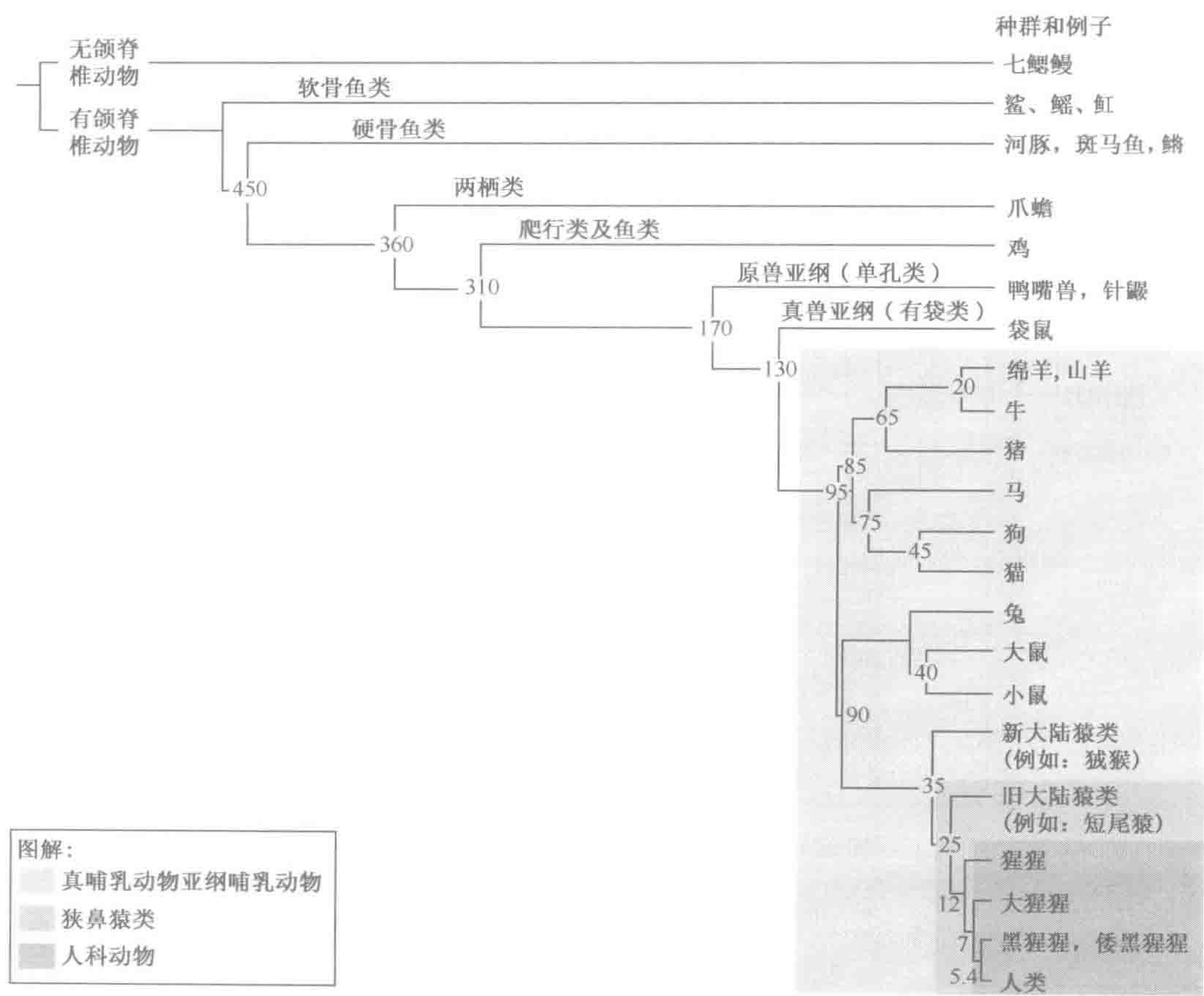


图 12.24 简化的脊椎动物系统发生图

节点处的数字表示估计的分化时间，以百万年为单位

大 *ORF2* 序列。

人-小鼠在散在重复序列含量方面的差异在很大程度上是由于人类基因组中较高的 SINE DNA 含量。这些 SINE 元件中无一来自共同的祖先（与 LINE 元件、LTR 元件以及 DNA 转座子不同，见 MGSC，2002；表 5 和 6）。在人类基因组中，仅一种 SINE，即 Alu 重复仍保持活跃，而在小鼠中却出现了四个 SINE 家族，均依赖于 LINE1 反转座机制。其中，B2、ID 以及 B4 重复源自 tRNA 基因的 cDNA 拷贝，而 B1 重复则与 Alu 重复相似，来源于 7SL RNA（图 12.25）。但 B1 与 Alu 重复在序列上已发生显著的歧化。

基因与蛋白质序列的歧化

约 80％的小鼠蛋白质似乎在人类基因组中具有严格的 1∶1 种间同源体，其序列一致性通常介于 70％～100％之间。然而，大约 10％左右的种间同源性人类与小鼠蛋白呈现更为极端的序列歧化（图 12.26）。许多分化最快的蛋白质据知作用于宿主的抵抗力/免疫，例如 MHC 基因，或者生殖过程。歧化迅速的生殖相关蛋白包括，转换蛋白 2

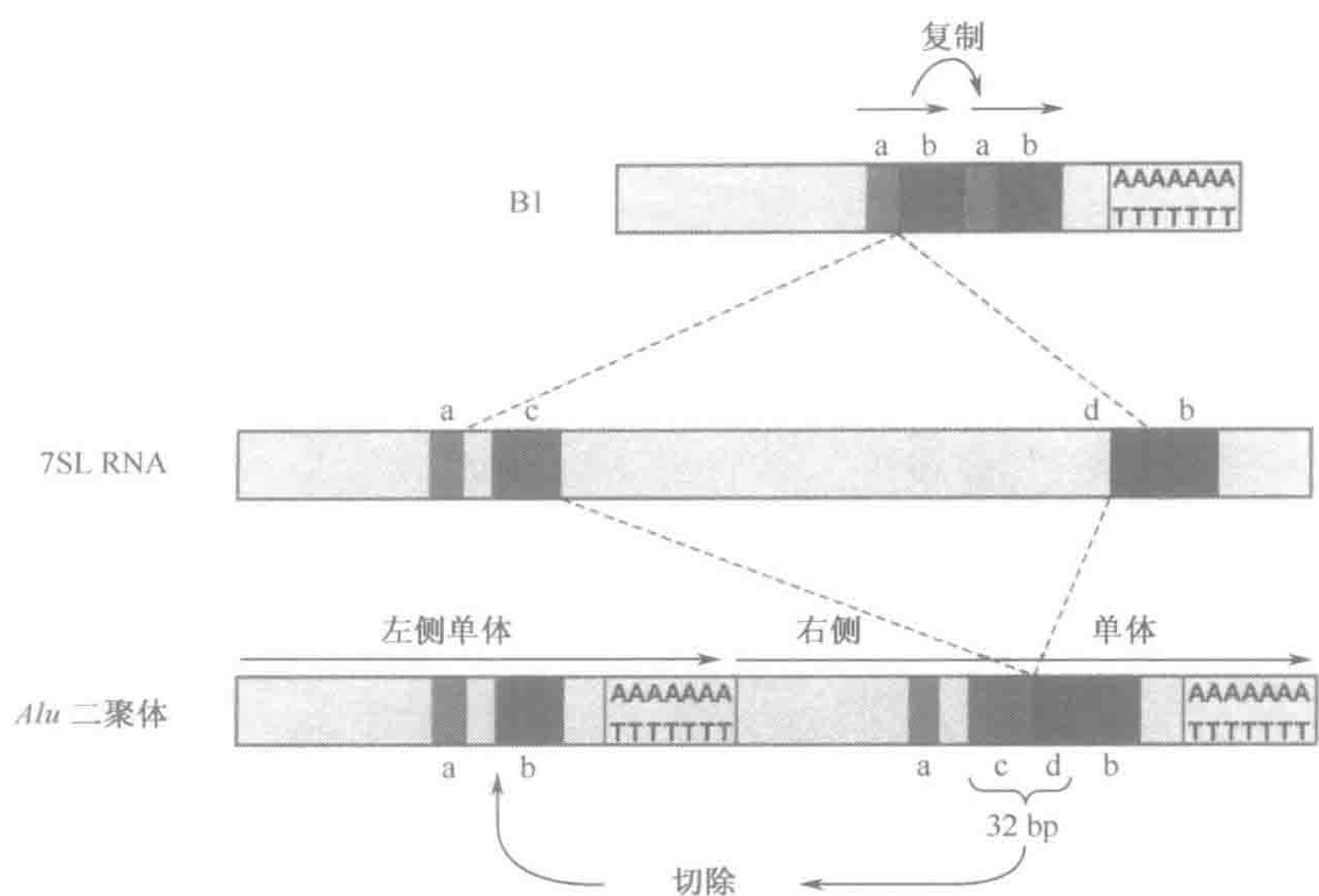


图 12.25 Alu 与 B1 重复进化自 7SL RNA 基因的加工后拷贝

Alu 重复序列与 7SL RNA 序列的两端之间广泛的同源性表明 7SL RNA 基因的一个多聚腺苷化拷贝通过反转座事件整合到了基因组的其他位置（其普通类型的机制见图 9.14）。在某些情况下，整合的拷贝可产生自身的 RNA 转录物。在一个很早的阶段，一个内部的片段（c 与 d 之间）丢失。随后，一个含有早期缺失（c + d）旁侧区域的 32 bp 中央片段又发生缺失，从而产生一个相关的重复单元。两类单元的融合产生了经典的 Alu 二聚体重复，其中左侧（5′）的单体缺乏上述的 32 bp 序列，而右侧（3′）的单体则含有该 32 bp 序列。**值得注意的是**，Alu 单体的多重拷贝亦见于人类基因组中。在小鼠中，似乎亦发生过一个类似的拷贝自 7SL RNA 基因的过程，形成一大段内部单元（a 与 b 之间）的丢失，随后又发生了旁侧区域（a + b）的串联重复。

（人类与小鼠氨基酸序列间差异度为 68%），透明带糖蛋白 2 和 3（差异度分别为 43% 和 33%），顶体酶（差异度为 38%），以及精子特异性精蛋白 P15 和 P2（差异度分别为 41%和 36%）。正向（达尔文）选择（对氨基酸替换进行选择以促进歧化）已在几种上述类型的蛋白质中被发现（例如 Swanson *et al.*,2001）。

基因数目的歧化

人类与小鼠基因的总数仍有待于确定，但估计大致相似。然而，许多人类与小鼠的蛋白质均属于曾在其中一个基因组中发生过不同扩张的基因家系，导致缺乏严格的1∶1对应关系。诚然，当一个基因家族中基因数目发生变化时，识别真正的种间同源基因将变得困难，尤其是存在较大序列歧化时。一个常设的例子就是主要组织相容性复合体（MHC），在后者中，种间同源基因无法从典型的 MHC 基因座（人类中的 HLA，小鼠中的 H-2）以及数目存在较大差异的非典型 MHC 基因座中被轻易地识别出来。

一个极端的例子就是嗅觉受体基因家族。小鼠具有大约 1200 个功能性基因，为人类功能性基因数目的三倍以上（Young *et al.*,2002），导致了小鼠与人类之间在气味探测方面的不同敏感度及潜能。另外还存在 25 个小鼠特异性基因簇，其中 14 个含有啮齿类生殖相关基因（MGSC，2002，表 16），5 个含有宿主抵抗力与免疫相关基因。由于

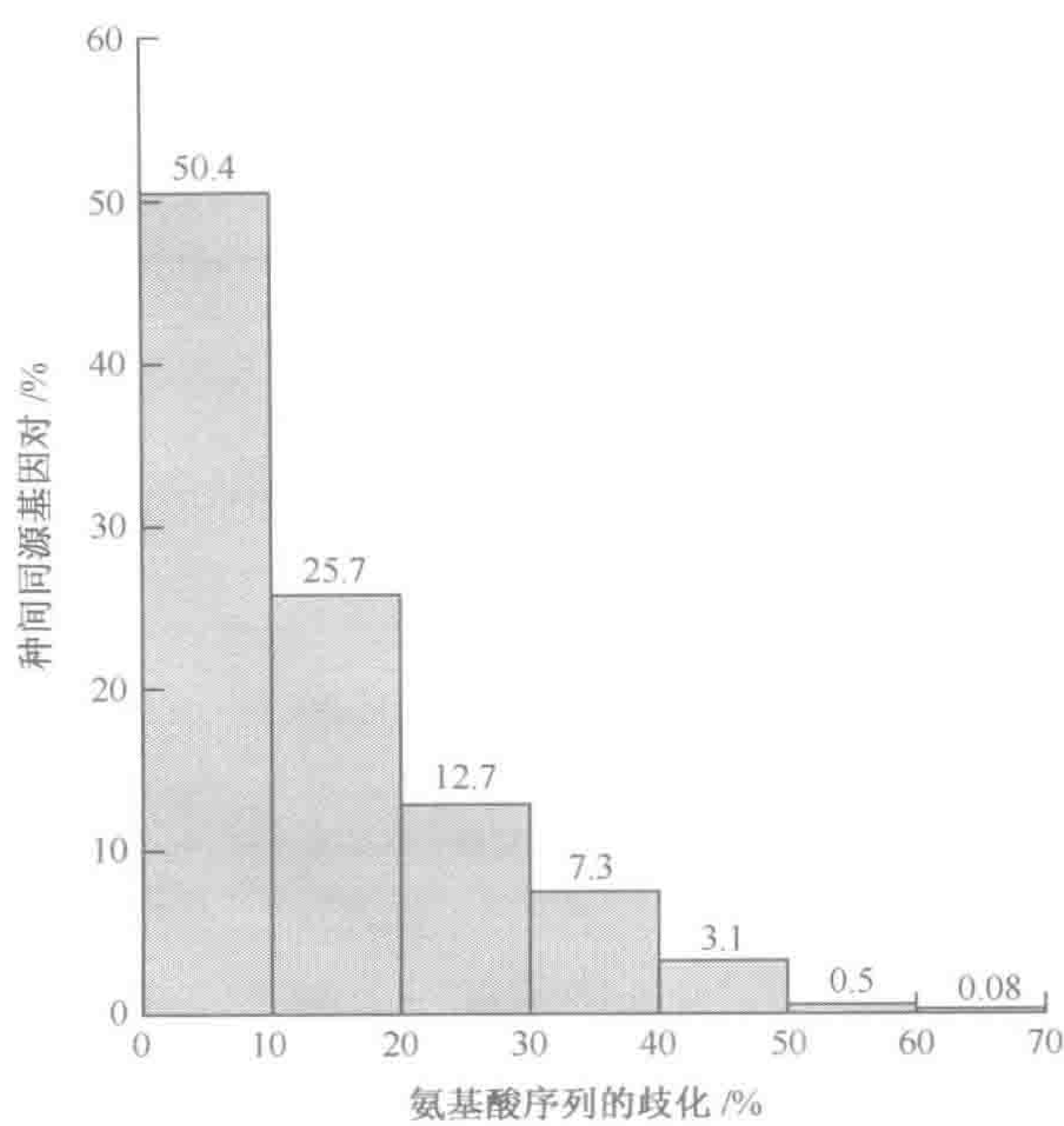


图 12.26　人-啮齿类蛋白质的歧化

一组包含 1880 对人-啮齿类种间同源基因的样本（最初由 Makalowski 和 Boguskin，Proc. Natl. Acad. Sci，USA 95，9402～9412. 1998 报道。）被根据氨基酸序列在种间同源体之间的差异水平分组。最高度保守的蛋白（如钙调蛋白、核糖体蛋白、组蛋白等）均涉及关键的细胞活动。演变最快的蛋白均涉及宿主抵抗力/免疫或生殖（正文）。

啮齿类与灵长类在生殖生理学上具有某些显著的差异，生殖性状可能是产生强大的进化压力的原因，对于更新的需求推动了基因家族的差异性扩张。

进化中的**基因丢失**（gene loss）亦可能导致基因数目不仅在成簇的基因家族而且在散在的基因中的差异。一些人类基因因此似乎在啮齿类种系中并没有种间同源体，尤其是对于主要假常染色体区域以及邻近的性染色体特异性区域内的基因来说（图 12.15）。发生于部分上述基因中的突变将导致疾病，诸如假常染色体区域基因 *SHOX*（Leri-Weill 综合征；Langer 肢中段发育不良）以及位于 Xp 上的 Kallman 综合征基因 *KAL*1（种间同源体被发现于秀丽新小杆线虫与果蝇中，但对应的基因似乎已从啮齿类种系中丢失）。

基因表达的歧化

人-小鼠在种间同源基因表达上的差别含有在 RNA 加工、启动子的选择性使用、X 染色体失活的模式以及遗传印记等方面存在差异的许多例子。此外，即使种间同源基因在蛋白质水平上高度保守，其表达模式的时空特征亦并非罕见地表现出显著的差别（Fougerousse *et al*.,2001）。

12.4.2　是什么使我们区别于我们最近的亲戚——大型猿类

在近一个半世纪之前赫胥黎（Thomas Huxley）即正确地辨认出黑猩猩和大猩猩是我们最近的亲戚。从那时起，进化遗传学家即开始与**三分法问题**（trichotomy problem）

进行斗争：两个物种中哪个才是我们最近的亲戚，抑或是人类、黑猩猩和大猩猩的种系同时发生了歧化（三分法）？大多数核苷酸测序数据（Satta *et al.*,2000）均支持人类与黑猩猩（包括 *Pan troglodytes*，即普通黑猩猩，以及 *Pan paniscus*，即矮黑猩猩或倭黑猩猩）之间的关系更为接近。人-黑猩猩进化支（＝群）亦得到了染色体断端分析的支持，后者显示在人类、普通黑猩猩以及矮黑猩猩的共同祖先中，一个长 0.1 Mb 的常染色体片段被转座到了 Y 染色体上。

框 12.5　常见后生动物种系发生群及术语释义

更多的信息请参见生命之树的网页（http://tolweb.org/tree/phylogeny.html）以及由进一步阅读所建议的其他系统发生学资料。简化的真核生物、后生动物以及脊椎动物系统发生见图 12.22-12.24。

羊膜动物（amniote）——具有羊膜的脊椎动物。包括爬行类、鸟类和哺乳类，但不包括鱼类和两栖类。

类人猿（anthropoid）——**即人科动物**（hominoid，请参阅）加上猴科。

海鞘类（ascidian）——海鞘（如玻璃海鞘），一种背囊动物（tunicate，请参阅）。

对称动物（bilaterian）——左右对称的后生动物，包括拥有左右对称性幼虫的棘皮类。它们具有三个胚层，因而目前与三胚层动物（triploblast，请参阅）同义。

狭鼻灵长类（catarrhine primate）＝**人科动物**（hominoid，请参阅）加上**旧大陆猿类**（old world monkey）。

头索动物（cephalochordate）[＝**鳃口动物**（branchiostome）＝**头索类**（lancet）]——不具有头骨或脊柱的脊索动物（如文昌鱼）。

脊索动物（chordate）——胚胎期出现脊索、背侧神经索以及咽部鳃囊的动物，包括**尾索类**（urochordate，请参阅）、**头索类**（cephalochordate，请参阅）以及**有头盖动物**（craniate，请参阅）。

刺细胞动物（cnidarian）[＝**腔肠动物**（coelenterate）]——水母、珊瑚虫、海葵（均呈放射状对称并具有两个胚层）。

有体腔动物（coelomate）——具有体腔或者充满液体、完全由中胚层包被的身体内腔的动物，与无体腔动物（如扁形虫）以及假体腔动物（如具有一个充满液体的身体内腔，但并未完全由中胚层所包被的线虫）相反。可分为两个群体，即**原口类**（protostome，请参阅）与**后口类**（deuterostome，请参阅）。

进化枝（clade）——包括由同一个最近共同祖先（单源性分类）所派生的所有生物。

有头类（craniate）——具有头骨的动物＝脊椎动物加上鱼类。

后口类（deuterostome）——来源于希腊语，意为**第二张口**（second mouth）＝**脊索动物**（chordate，请参阅）加上**棘皮动物**（echinoderms，请参阅）。胚孔（原始消化腔向外的开口）位于胚体后部并成为肛门的动物。之后，嘴巴开口于肛门的对侧。与**原口类**（protostome）相对应（请参阅）。

双胚层动物（diploblast）——仅具有两个胚层的后生动物，即外胚层和内胚层＝**刺细胞动物**（cnidarian，请参阅）＋栉水母类（栉水母）。它们呈放射状对称，因此该种群通常被归类为**辐射动物**（radiata，请参阅）。

棘皮类动物（echinoderm，如海鞘，海星等）为放射状对称的海洋动物，但由于具有三个胚层，它们为**三胚层动物**（triploblast），因此被归类为**对称动物**（bilaterian，请参阅）。

真哺乳亚纲动物（eutherian mammal）——胎生哺乳动物，与**单孔类**（monotreme，请参阅）以及**后兽亚纲**（metatherian mammal，请参阅）动物相对。

框 12.5　常见后生动物种系发生群及术语释义（续）

有颌类动物（gnathostome）——有颌的脊椎动物，占脊椎动物的大多数，但七鳃鳗为无下颌的脊椎动物。

八目鳗（hagfish）——脊椎动物的近亲，但属于无脊椎动物，因为其脊索并不能转变为一条脊柱。

原始人类（hominid）——人类以及类似于人的祖先。

人科动物（hominoid）＝人类加上大猿类（普通黑猩猩、倭黑猩猩、大猩猩、猩猩）以及小猿类（长臂猿）。类人猿比较（anthropoid，请参阅）。

后兽亚纲哺乳动物（metatherian mammal）（＝有袋类动物）。

单孔类（monotreme）——卵生哺乳动物。

新大陆猿类（new world monkey，阔鼻类）——局限于墨西哥南部、中、南美洲热点雨林中的猴类。

旧大陆猿类（old world monkey，猕猴科）——发现于南亚、东亚、中东以及非洲广泛环境中的猴类。

门（phylum）——具有共同身体结构的一组物种。

阔鼻灵长类（platyrrhine primate）＝**新大陆猿类**（New world monkey，请参阅）。

原兽亚纲哺乳动物（prototherian mammal）（＝**单孔类**，monotreme，请参阅）。

原口类（protostome）——来源于希腊语，意为第一张口。嘴巴起源于胚孔（原始消化腔向外的开口）的动物。之后，肛门开口于嘴巴的对侧。包括软体动物，环节动物以及节肢动物。与**后口类**（deuterostome）相对应（请参阅）。

辐射动物（radiata）——放射状对称动物，仅具有两个胚层＝双胚层动物（diploblast）（请参阅）。

分类（taxon）——在分类系统的任何一级被归为一群的生物。

硬骨鱼（teleost fish）——**多骨鱼**（bony fish）中的一大类（与具有软骨性骨骼的鱼类如鲨鱼、鳐以及鳐等相对）。以完全活动的上腭，刺鳍以及鱼鳔为特征。

三胚层动物（triploblast）——具有三个胚层的后生动物。它们呈左右对称，因此与**对称动物**（bilaterian）同义。

背囊动物（tunicate）——海鞘以及樽海鞘（如玻璃海鞘，即海鞘）。

尾索动物（urochordate）——脊索动物（chordate，请参阅），具有一个仅限于尾部区域的脊索（＝**背囊动物**（tunicate））。

人-黑猩猩以及人-大猩猩种系的歧化分别发生于大约 500 万～600 万年以及大约 700 万年之前，导致了一系列遗传学差异（Gagneux and Varki，2001；Olson and Varki，2003）。向不同物种的歧化（物种形成）可能最初由小型的细胞遗传学差异，如倒置（在哺乳动物中可抑制重组），以及/或调节配子形成或早期胚胎发育的关键基因的突变所推动。

基因组的构成

经典的细胞遗传学比较所强调的是人科动物（人＋大型猿类）染色体显带模式非常强烈的保守（Yunis and Prakash，1982）。大型的结构差异似乎有限，包括若干臂内及臂间倒置，近期由两条染色体融合而成的人类 2 号染色体，以及与人类 5 号及 17 号相对应的两条大猩猩染色体之间的相互易位等（图 12.27）。着丝粒序列呈现非常迅速的进化，虽然一段 α 卫星序列在所有人类及大型猿类的染色体中均得到了保留，但仍存在

很可能是由个别种系中重复序列的共同进化所形成的显著序列差异。最近的运用多色荧光原位杂交（FISH）的比较定位研究（图 12.28）已证实强烈的同线性保守，并推测出一个人科动物的原始核型（Muller and Wienberg，2001）。

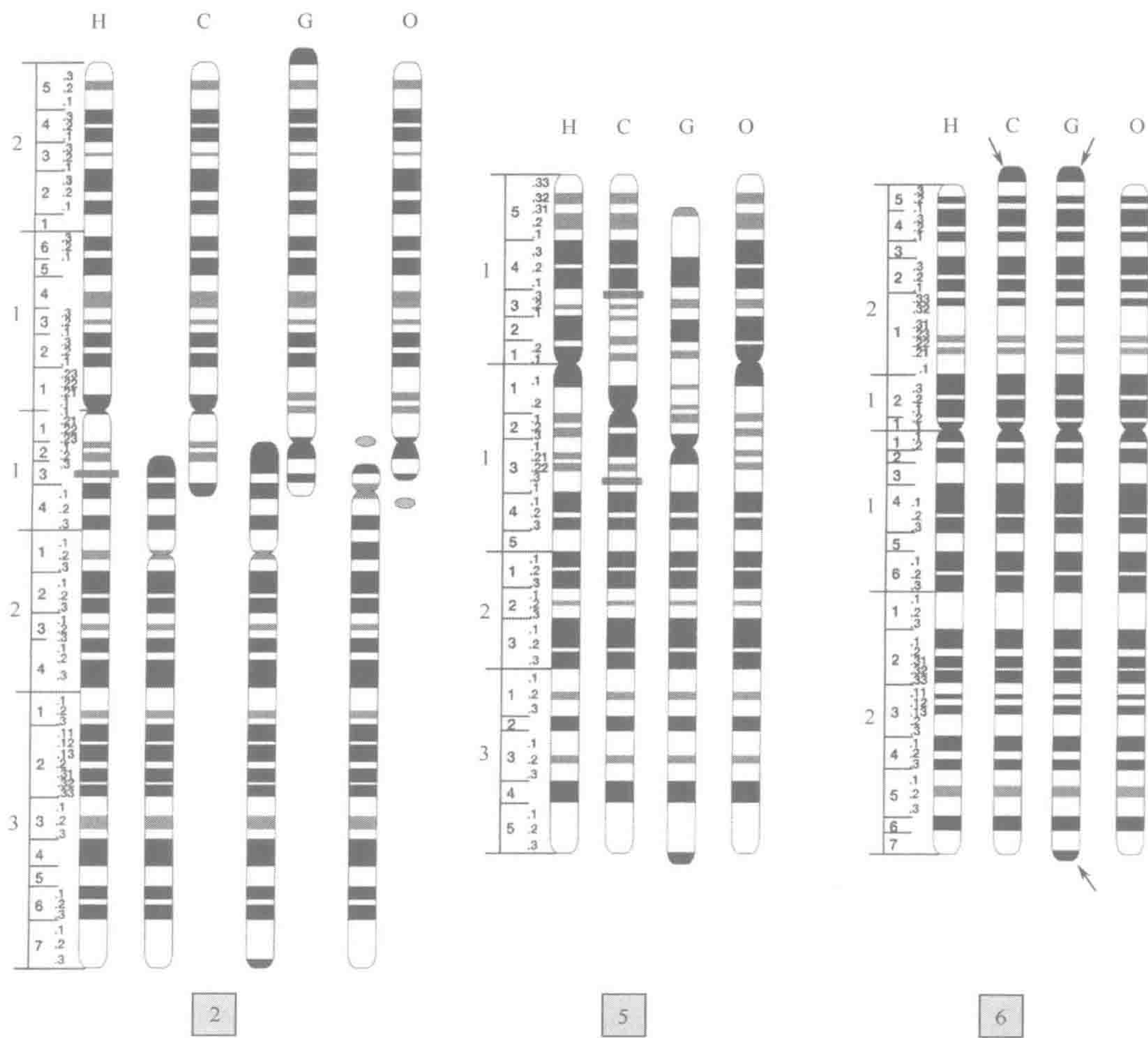


图 12.27　人类染色体的带型与类人猿非常相似

示意图源自 1000 条带的人类（H）、黑猩猩（C）、大猩猩（G）、猩猩（O）的晚前期染色体。人类的 2 号染色体产生自两条灵长类染色体的融合。人类的 6 号染色体则与其种间同源体极其相似——唯一容易辨认的差别在于黑猩猩同源体的短臂，以及大猩猩同源体的两条臂上均具有额外的端粒区异染色质。人类 5 号染色体的种间同源体之间具有显著的差异。黑猩猩的同源体曾经历一次对侧倒置（断端位置对应于 5p13 与 5q13），大猩猩的同源体则曾经历过一次与人类 17 号相应的染色体的相互易位。经 Yunis 和 Prakash(1982). Science 215，1525～1530. ©1982，美国科学促进会允许重印。

序列采样分析提示大约 95％的黑猩猩基因组能够被直接与相应的人类序列比对，这些匹配区域内的序列变异平均为 1.2％；不匹配的 5％则是缘于缺失/插入（Olson and Varki，2003）。尽管存在非常高的序列一致性（非编码 DNA 以及编码 DNA），一些人类特异性的 Alu 亚家族（Sb1，Sb2）以及 LTR 反转座子仍可识别。

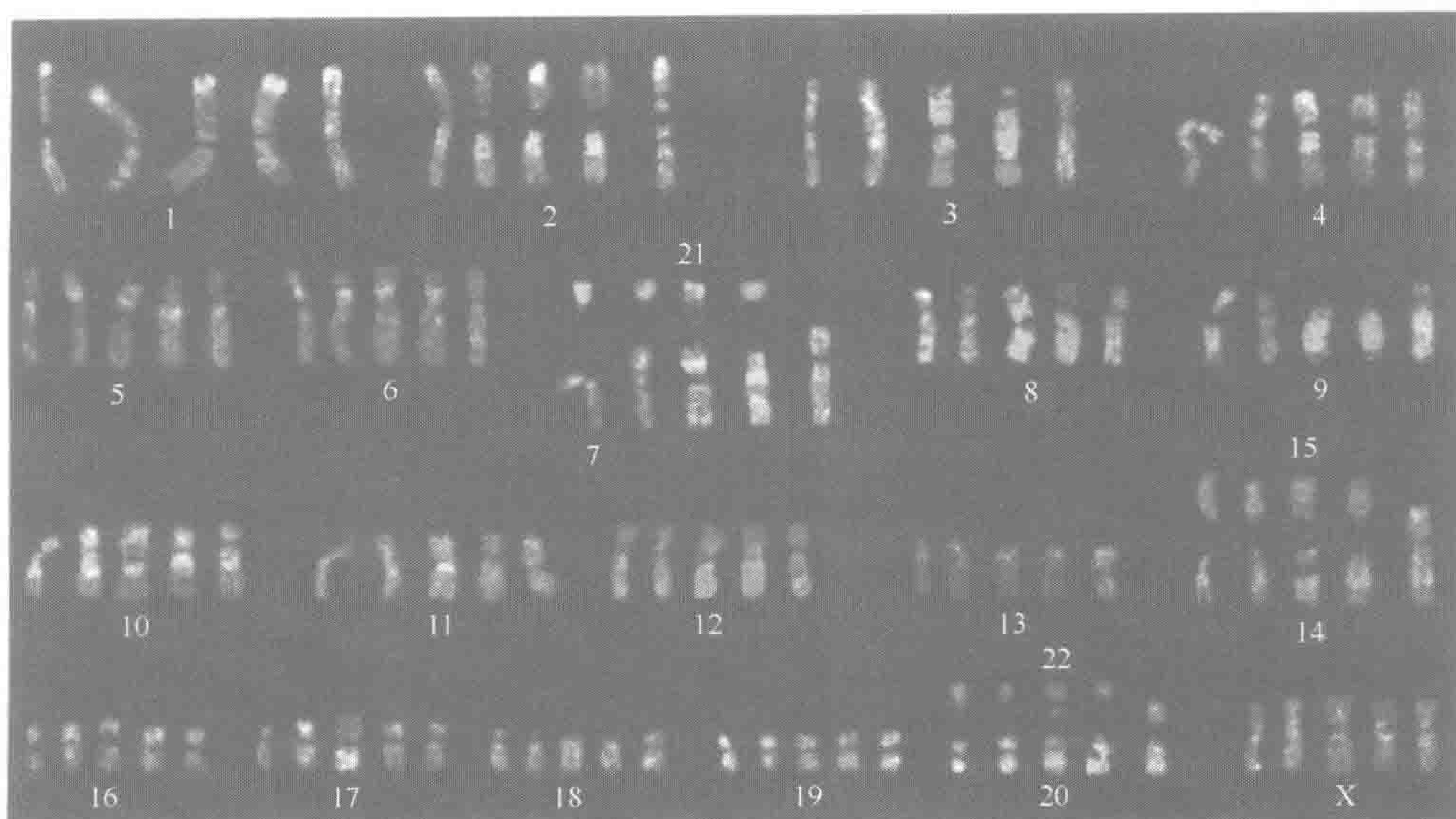


图 12.28 对灵长类染色体进行条码化以显示其结构上的差异

跨物种的彩色显带（Rx-FISH）谱显示了种间同源的灵长类染色体与人类 1～22 号以及 X 染色体（依据人类种间同源体编号）的比对，自左向右依次为：人、黑猩猩、大猩猩、猩猩及短尾猿。为了突出对比，人类或短尾猿中的单条染色体 2p/2q、7/21、14/15 以及 20/22 与类人猿的种间同源体放在一起。这类分析提示了人类及类人猿的一种初步的原始核型。经柏林 Springer-Verlag 出版社允许，复制于 Muller and Wienberg(2002). Hum. Genet. 109，85～94。

基因的差异

当种间同源性人类与黑猩猩序列被比较时，编码区 DNA 通常显示高于 99％的序列一致性。在某些时候，某些人类基因的特殊等位基因与黑猩猩的种间同源体具有较其他人类等位基因更近的关系。例如，在人类 HLA-DRβ 位点，等位基因 *HLA-DRB1*0302* 和 *HLA-DRB1*0701* 与同源性黑猩猩基因 *Patr-DRB* 在序列上的关系明显较二者之间更近（图 12.29）。这一现象与这些歧化的等位基因相对古老的起源，即早于人-黑猩猩的分化相一致。

灵长类特异性的基因已经被发现，其中包括一些具有最近进化起源以及一些似乎曾经历对氨基酸替换的正向选择者（Courseaux and Nahon，2001；Johnson *et al*.,2001）。人类特异性的基因可能很稀少，尽管基因复制与基因丢失——尤其是在成簇的基因家族中——可能造成人类与大型猿类之间基因数目的差异。然而，目前并未发现证据显示人类与黑猩猩的基因之间存在功能上的差异：唯一已知的生物化学差异就是在人类中，一种特殊唾液酸，即表达（以发育学上的调节以及组织特异性的方式）于黑猩猩、倭黑猩猩以及其他大型猿类中的 N-糖基乙二醇神经氨酸的全面减少。这种差异来源于 CMP N-乙酰神经氨酸基因内的一个移码突变，后者于大约两百万年前发生于人类种系中，恰好在人脑的迅速扩张开始之前（Chou *et al*.,2002）。

权头结构域基因 *FOXP2* 是首个被与人类所独有的说话与语言特征相关联的基因。

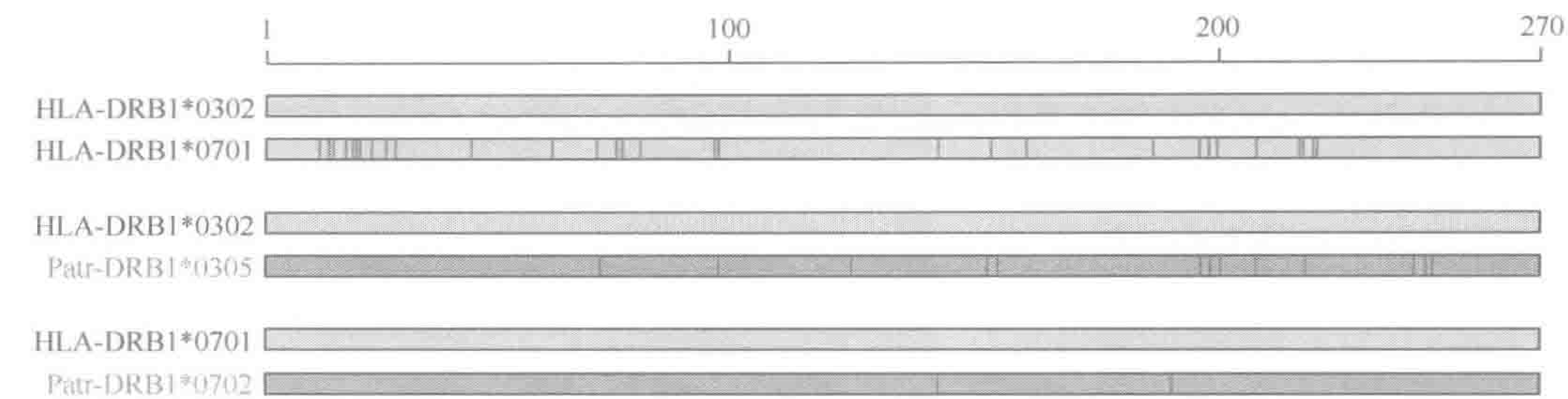


图 12.29　一些人类等位基因之间呈现较分别与种间同源的黑猩猩基因相比时更大的序列差异

在总共 270 个氨基酸位置中，*HLA-DRB1* * 0302* 与 *HLA-DRB1* * 0701* 等位序列共显示出 31 处差异（13%）。将其中任意一条分别与黑猩猩的种间同源基因座上的序列（*Patr-DRB1*）进行比较则发现了更为接近的人-黑猩猩基因对，如 *HLA-DRB1* * 0701* 与 *Patr-DRB1* * 0702*（在 270 个氨基酸中仅两个不同）。这提示某些现存的 HLA 等位序列早于人-黑猩猩的分离出现。经《科学美国人》出版公司允许，据 Klein 等（1993）. Scientific American 269，675～680 重绘。

它在进化中得到了很好的保留，在人类与小鼠之间仅显示三种氨基酸的替换。然而，这些替换中的两种为人类所独有（大型猿类具有与小鼠相同的氨基酸），并推测在人类近期的进化中曾受到正向选择（Enard *et al*.,2002a）。可能的情况是，它们能够影响一个人控制说话所需要的口面部运动能力。

有两种系统方法最近被推荐用于寻找人类独特性状的分子学基础。第一种是通过基于微阵列芯片的 RNA 分析以及基于双向电泳的蛋白质分析对人类与灵长类的基因表达进行全基因组范围的比较。初步的数据显示，物种特异性基因表达模式在大脑部位相当突出（Enard *et al*.,2002b）。第二种策略是在人类基因组序列中筛查曾在人类种系演化中历经强烈选择的区域。对于一种具有优势的新等位基因的强力选择将可能引起一次**选择性清扫**（selective sweep，随着新突变的频度升高，邻近的染色体区域亦将会被清扫以固定下来），产生核苷酸多样性极低的区域。各种人类基因被发现于这类区域中，并为找出现代人类与黑猩猩之间的遗传差异提供了目标（Diller *et al*.,2002）。

12.5　人类种群的进化

渴望了解人类种群是如何进化的超越了我们对过去的简单好奇心。它亦将发掘出从分子水平上适应环境的证据，并有助于解读和预测连锁不平衡，进而影响对于疾病的关联分析。

在 DNA 分析产生影响之前，对人类种群进化的研究依赖于考古及化石的分析。非洲的原始人类化石证据开始自约 4 百万年前的上新世早期（地质学年代 530 万至 180 万年前），其代表为来自埃塞俄比亚与坦桑尼亚的南方古猿（*Australopithecus*）。直立人（*Homo erectus*）出现于一百多万年之前的更新世（地质学年代 180～80 万年前），并产生了我们本身的物种。一些研究者则支持来源于直立人的另一个物种海德堡人（*Homo heidelbergensis*）依次产生了智人（*Homo sapiens*）与尼安德特人（*Homo neanderthalensis*）的可能性（Stringer，2002）。具有现代解剖学特征的人类（*Homo sapiens*

sapiens）开始出现于12万至10万年之前，并与尼安德特人并存，直至后者于大约3万年之前灭绝。

DNA变异为重构人类种群的近期历史提供了另一种途径。不同的DNA标记可被采用。线粒体DNA与非重组性Y染色体序列备受欢迎，因为与其他标记不同，它们不易发生重组，并可能通过母系（mtDNA）或父系（Y染色体）直接进行种系追溯。当然，它们并不足以代表整个基因组，因此常染色体以及X连锁的序列亦被用作研究。

除了对现代人类不同种群中的DNA变异进行测定之外，DNA亦可回收自考古学的骨骼标本。尽管长度较短，经回收的DNA序列仍可能经PCR扩增，之后通过测序或基因型鉴定进行分析。降解过程造成了时间的局限，但通过小心操作，仍可能从早至5万年之前的标本中提取出**古DNA**（ancient DNA）。回收自3～10万年前尼安德特人标本的DNA具有与现代人类明显不同的mtDNA序列：约3倍于人类个体之间的平均差异，但同时仅为人类与黑猩猩之间的平均差异的一半（Krings *et al.*,1997）。这一数据，虽然难免有限，提示尼安德特人并未为现代人类贡献任何mtDNA就绝迹了，而产生尼安德特人与现代人类的种系分离于大约50万年之前。

12.5.1　遗传学证据提示现代人类具有一个自非洲人群的最近起源

当应用于现代人类种群时，DNA变异研究清晰地提示现代人类具有一个自非洲人群的最近起源（Excoffier，2002；Templeton，2002），并与见于非洲人群之间较大的遗传多样性相一致（见下一节）。一个较早的假说，即**最近非洲起源模型**［recent African origin(RAO)model，又称单地区假说］认为人类物种进化于一个较小的非洲群体，并随后于大约10万～15万年前（约在具有现代解剖学特征的人类出现时。图12.30）迁移至整个世界，并取代了原先的原始人类。

图12.30　由最近非洲起源模型（单地区假说）所推测的智人自东非向外的扩散

数字为根据化石及考古学证据所估算的现代人到达各地的时间（迄今的年数）。由遗传学研究所提示的、支持最近非洲起源模型的人类迁移出非洲的时间范围介于50 000～200 000年前之间。经柏林Springer-Verlag出版社允许，据Klein and Takahata(2002). Where do we Come From? The molecular evidence for human descent重绘。

RAO 模型最初建立在对人类 mtDNA 单体型变异的估算上。mtDNA 的母系遗传以及重组的缺乏意味着**溯祖分析**（coalescence analysis，框 12.6）能够方便地被应用于估算共同的祖先，将被研究的 DNA 序列转移至全部采样个体的年代。利用这种办法，所有现代人类的 mtDNA 能够被追溯至一个个体，即‘线粒体夏娃’上。分析显示这个个体存在于大约 10～15 万年（5000～7000 代左右）之前，并且数据有力地提示她生活在东非（见下文）。当然，线粒体夏娃并非是当时生活在这个行星上的唯一的人：大约有 1 万个个体生活在那个时候，但与夏娃不同的是，他们的 mtDNA 序列并未被传递至现有的人类种群（框 12.6）。

由于我们的直接祖先，即直立人，据知已于一百多万年之前扩散出非洲，RAO 模型仍具有争议性。这一模型因此需要假定仅非洲的直立人产生了现代的人类。尽管最初被建立在不正确的估算上，RAO 模型得到了更为可靠的最新数据的支持。由图 12.31 所示的基于 mtDNA 的系统发生学提供了一个常规分析的例子。其关键在于最深的分支（最早从根上分出者，因而在进化上也更早）毫无例外地指向发现于非洲种群中的变异型，强烈地显示线粒体夏娃生活在非洲。重建进化树的方法（如近邻法，如图 12.31 所示）可以对所有现存的 mtDNA 序列并合为一条单一的原始 mtDNA 序列所花费的时间进行统计学估算。

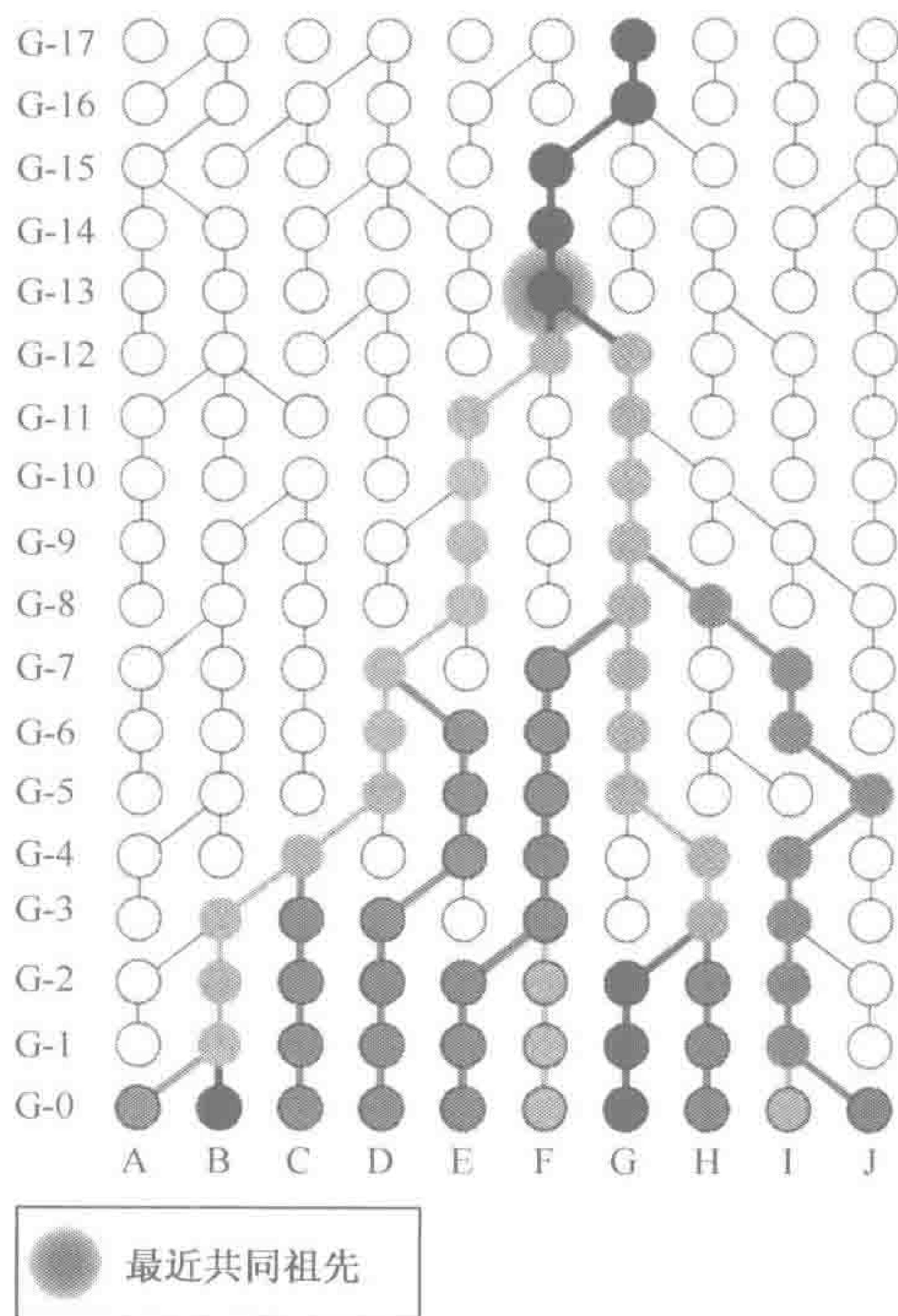
框 12.6　溯祖分析

在进化遗传学中，**并合**（coalescence）与歧化相对立。在正常的进化方向上，从过去到现在，基因与 DNA 序列分化于一条共同的原始序列。溯祖分析的意义在于从现存某些基因座上的多样性出发来试图对进化过程进行回溯。诚然，如果仅对父系或母系进行追踪，这将容易得多，因此对于 mtDNA（母系）与非重组性 Y 染色体序列（父系）来说，溯祖分析相对较为直截了当，因为没有导致混淆的重组。

在收集到不同群体的标本之后，溯祖分析即试图来确定可将所有存在于标本中的差异并合入一条存在于**最近共同祖先**（most recent common ancestor，MRCA）中的原始序列的时间点。并合概念也暗示了血统的不平等性：所有的现代人类及其基因起源于携带原始基因的祖先个体，但并非所有存在过的人均拥有后裔。不同的世系因此可通过不同的最近共同祖先（见图）相联系。最终，出现于现代人类某个基因位点上的所有遗传变异皆能够被追溯至某一个体，诸如对 mtDNA 来说的线粒体夏娃。

重要的是，需要了解不同的位点可能将导致向不同最近共同祖先的并合。在现今地球上存在的三十亿条左右的人类 Y 染色体并合于一个生活在过去的个体，即‘Y 染色体亚当’，所携带的一条 Y 染色体上。他可能并未与线粒体夏娃生活在同一时期，因为他所携带的 Y 染色体的进化途径与夏娃的线粒体 DNA 并不相同。除了某些特殊位点，重组造成了对于常染色体序列分析的困难。然而，重组确实又倾向于发生在某些基因组区域而非其他，我们的常染色体基因组似乎属于由所谓**单体型区块**（haplotype block，通常长 5 kb～ 200 kb 的片段，含有 3～7 种代表了现代人群大部分的遗传变异；Paabo，2003）镶嵌而成。单体型区块在非洲人中趋向于更短，因此人种范围的单体型区块平均长度可能在 100 kb 左右。每个常染色体单体型区块都可能具有自己的进化史，因此，除线粒体夏娃与 Y 染色体亚当之外，数以千计的其他最近共同祖先亦参与了现代人类遗传性 DNA 的形成。

框 12.6　溯祖分析



溯祖分析寻求对基因或 DNA 序列世系进行回溯，直至它们并合于一个个体

现存人群中特定基因或 DNA 序列的不同世系（世代 G-0，底部的一排）能够最终回溯至存在于既往世代（G-1 至 G-17）中的不同最近共同祖先。E 是最早的世系分支，分化于八个世代之前（G-8，自世系 F-H）。全部世系并合于一个存在于十三个世代之前（G-13）的最近共同祖先。空白圆圈表示所研究的基因/DNA 序列未被传递至目前世代的情况。

另一种**多区域进化**（multiregional evolution）的模型得到了部分古生物学家的支持。在这里，现代人类被推想为逐渐而同时出现于散在于不同大陆的直立人群体中，不同的种群之间又存在显著的基因漂流。对于单体型树的最新统计学分析亦对一个简单的 RAO 模型提出了争议。在证实非洲在现代基因库形成中所扮演的主要角色的同时，Templeton(2002) 提出人类曾不止一次从非洲扩散出去，并在局部区域中发生过异种繁殖（各种模型见 Excoffier，2002）。

12.5.2　人类的遗传多样性偏低且大多由于群体内而非群体间的差异

研究一致显示，与包括小鼠和猿类在内的其他物种相比，人类的遗传多样性偏低，提示人类种系曾在进化的最近阶段经历一个**人口瓶颈**（population bottleneck）。变异的模式并不一致：非非洲裔人群通常呈现部分见于非洲人群的遗传变异。

一个惊人的例子是见于胰岛素微卫星位点上的变异，后者已被与糖尿病以及其他表型的易感性相关联。最近一项对三个非洲人群以及三个非非洲人群的研究共发现了 22 个高度歧化的相关等位序列谱系。在现存的人群中并未发现这些谱系之间的结构性过渡

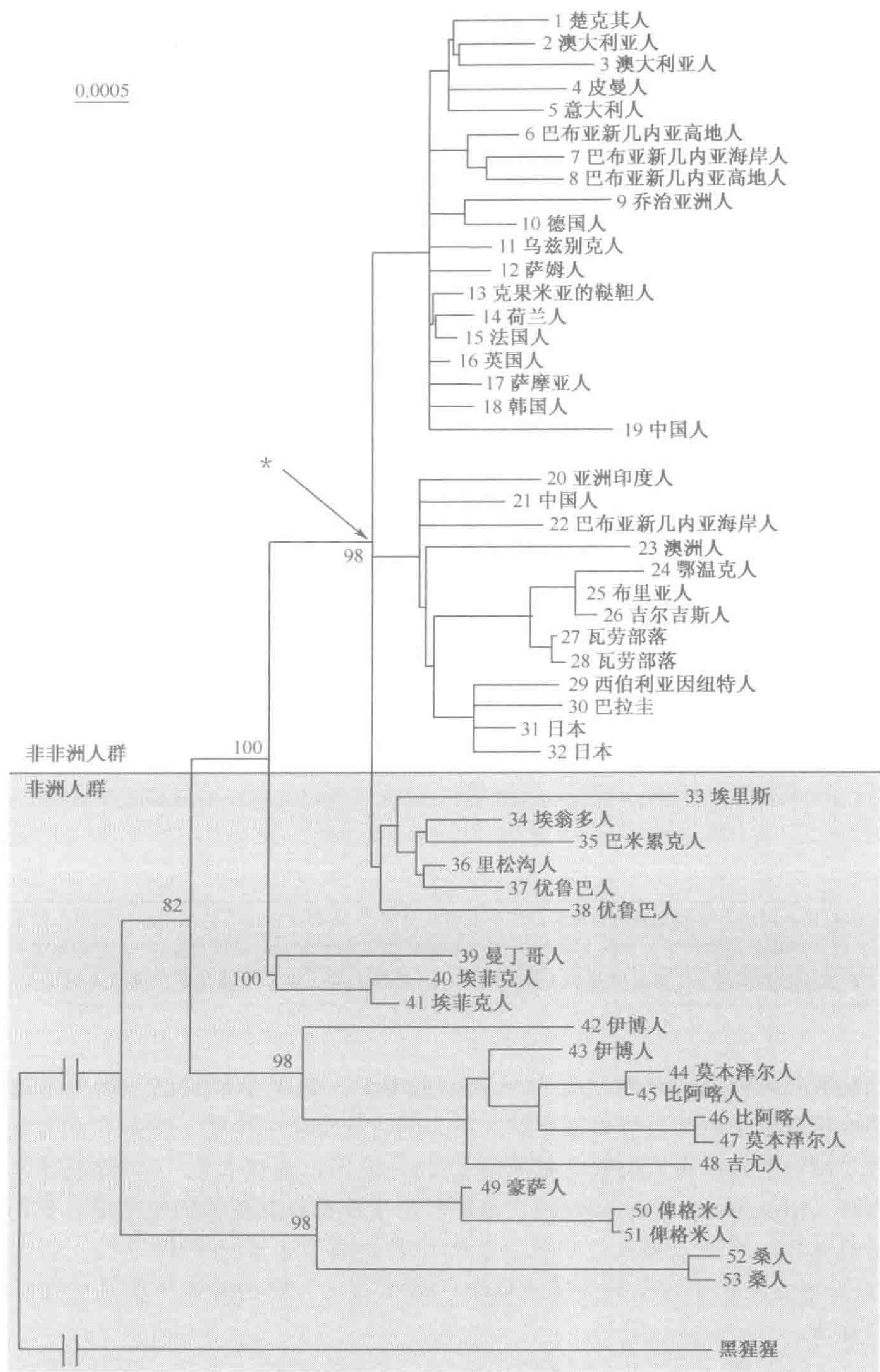


图 12.31　mtDNA 的系统发生学提示现代人类具有一个最近的非洲起源

由邻接法所构建的基于来自 53 名个体（人群起源如右侧标记，黑猩猩被作为外部参考物种）近全长 mtDNA 序列（D 环除外）的系统发生。数据为重复 1000 次的自举检验（自举值如节点上的数字所示）。具有非洲血统的个体见于中央水平线下方，其他人群则在上方。由星号所标注的节点表示同时含有了非洲以及其他人群个体的一个最晚进化分支的最近共同祖先。这个分析提示现代人类于大约五万年前起源于非洲，略较许多其他类似的分析结果为近。经 Nature 出版集团允许，复制于 Ingman 等．(2000). Nature 408，708～713。

物，提示在所有人类的世系中曾存在一个瓶颈（Stead and Jeffreys，2002）。非洲与非非洲人群之间的差异非常之大，所有 22 个谱系均在非洲被发现，但仅 3 个谱系发现于非非洲人群中，显示一个走出非洲的共同起源（表 12.6）。

表 12.6 胰岛素小卫星序列上非洲人群较非非洲人群更大的遗传多样性

种系	英国	非非洲人群 哈萨克斯坦	日本	象牙海岸	非洲人群 津巴布韦	肯尼亚
I	71.6	85.0	94.9	19.2	18.8	15.5
ⅢA	23.0	11.3	4.2	5.8	1.4	7.1
ⅢB	5.4	3.8	0.8	3.2	1.4	3.6
F	—	—	—	—	0.7	—
G	—	—	—	1.3	0.7	—
H	—	—	—	—	1.4	3.6
J	—	—	—	9.6	8.0	9.5
K	—	—	—	20.5	17.4	20.2
L	—	—	—	8.3	5.1	3.6
M	—	—	—	3.2	0.7	3.6
N	—	—	—	2.6	5.8	—
O	—	—	—	2.6	—	—
P	—	—	—	—	2.9	2.4
Q	—	—	—	1.9	5.8	9.5
S	—	—	—	4.5	.7	3.6
T	—	—	—	.6	4.3	1.2
U	—	—	—	—	1.4	—
V	—	—	—	1.3	1.4	4.8
W	—	—	—	12.8	13.0	9.5
X	—	—	—	1.9	4.3	2.4
Y	—	—	—	0.6	3.6	—
Z	—	—	—	—	0.7	—

对 INS VNTR 基因座上不同重复的分布的高精度分析共发现了 22 种高度歧化的等位序列世系。世系的频率由百分比表示。仅三种世系被发现于非非洲人群中（-符号代表不存在与世系关联的等位基因）。与此形成强烈反差的是，全部 22 种世系均被发现于人群之间差异大得多的非洲。其他人群中世系 I 的过多出现可能是缘于正向选择。经芝加哥大学出版社允许，根据 Stead 和 Jeffreys（2002）. Am. J. Hum. Genet. 71，1273～1284 重绘。

尽管在不同人群之间存在如表 12.6 所示的差异，遗传学研究已一致显示绝大多数人类的遗传学差异来源于人群内部而非人群之间。这类研究中的一次最全面者曾发现人群内个体之间的差异可占人类遗传差异的 93%～95%，而仅 3%～5%的差异来源于主要人群之间（Rosenberg *et al.*,2002）。尽管每个个体都有其独特的生活史，我们的生活史是如此的重合，以至于将人们划分为生物学种类将具有相当的误导性。因此，认为"种族"是可由生物学来定义的分散的概念已毫无意义，因为并无可能从生物学角度来定义这些类型。

由 Rosenberg 等（2002）所进行的研究是一个新的开端，可在不依赖被研究对象的地理起源这一前提信息的情况下确立人群的遗传学结构。在缺乏这种信息的情况下，五个主要的遗传簇被发现，分别对应于五个主要的地理区域：亚撒哈拉非洲；撒哈拉非洲加上欧亚地区（=欧洲+中东+中/南亚）；东亚；美洲（美洲印第安人）以及大洋洲等。因此，在地理与遗传学分区之间存在很好的一致性。即便如此，如前所述，在上述

任一地理区域中来源于两个个体的两个基因之间平均仅比来源于同一区域中不同个体的两个基因更相似 4%。

（李　岭　译）

进一步阅读

Carroll SB (2003) Genetics and the making of *Homo sapiens*. *Nature* **422**, 849–857.

Felsenstein J (2003) *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, MA.

Graur D, Li WH (2000) *Fundamentals of Molecular Evolution*. 2nd Edn. Sinauer, Sunderland, MA.

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. *Nature Rev. Genet.* **4**, 275–284.

Klein J, Takahata N (2002) *Where do we Come From? The molecular evidence for human descent*. Springer-Verlag, Berlin/Heidelberg.

Meyer A, Van de Peer Y (2003) *Genome Evolution: gene and genome duplications and the origin of novel gene functions*, Kluwer, Dordrecht.

Nei M, Kumar S (2000) *Molecular Evolution and Phylogenetics*. Oxford University Press, Oxford.

Olson S (2002) *Mapping Human History: discovering the past through our genes*. Houghton Mifflin Co., Boston.

Relethford JH (2001) *Genetics and the Search for Modern Human Origins*. Wiley, New York.

Smith JM (1999) *Evolutionary Genetics*. 2nd Edn. Oxford University Press, Oxford.

参考文献

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of *en bloc* duplication in vertebrate genomes. *Nature Genet.* **31**, 100–105.

Bailey JA, Gu Z, Clark RA *et al.* (2002) Recent segmental duplications in the human genome. *Science* **297**, 1003–1007.

Baxendale S, Abdulla S, Elgar G *et al.* (1995) Comparative sequence analysis of the human and pufferfish Huntingtons disease genes. *Nature Genet.* **10**, 67–75.

Bonen L, Vogel J (2001) The ins and outs of group II introns. *Trends Genet.* **17**, 322–331.

Brown JR (2003) Ancient horizontal gene transfer. *Nature Rev. Genet.* **4**, 121–132.

Burmester T, Weich B, Reinhardt S, Hankeln T (2000) A vertebrate globin expressed in the brain. *Nature* **407**, 520–523.

Carrel L, Cottle AA, Goglin KC, Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proc. Natl Acad. Sci. USA* **96**, 14440–14444.

Charchar FJ, Svartman M, El-Mogharbel N (2003) Complex events in the evolution of the human pseudoautosomal region 2 (PAR2). *Genome Res.* **13**, 281–286.

Chou H-H, Hayakawa T, Diaz S *et al.* (2002) Inactivation of CMP-*N*-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc. Natl Acad. Sci. USA* **99**, 11736–11741.

Ciccodicola A, D'Esposito M, Esposito T (2000) Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **9**, 395–401.

Courseaux A, Nahon J-L (2001) Birth of two chimeric genes in the *Hominidae* lineage. *Science* **291**, 1293–1297.

Diller KC, Gilbert WA, Kocher TD (2002) Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. *Mol. Biol. Evol.* **19**, 2342–2345.

Doolittle WF (1998) You are what you eat: a gene transfer rachet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* **14**, 307–311.

Dubchak I, Brudno M, Loots GG *et al.* (2000) Active conservation of noncoding sequences revealed by 3-way species comparisons. *Genome Res.* **10**, 1304–1306.

Ellis N. (1998) The war of the sex chromosomes. *Nature Genet.* **20**, 9–10.

Enard W, Przeworski M, Fisher SE *et al.* (2002a) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869–872.

Enard W, Khaitovich P, Klose J *et al.* (2002b) Intra- and inter-specific variation in primate gene expression patterns. *Science* **296**, 340–343.

Excoffier L (2002) Human demographic history: refining the African model. *Curr. Opin. Genet. Dev.* **12**, 675–682.

Fougerousse F, Bullen P, Herasse M *et al.* (2000) Human-mouse differences in the embryonic expression patterns of developmental control genes and disease genes. *Hum. Mol. Genet.* **9**, 165–173.

Gagneux P, Varki A (2001) Genetic differences between humans and great apes. *Mol. Phylogenet. Evol.* **18**, 2–13.

Gianfrancesco F, Sanges R, Esposito T (2001) Differential divergence of three human pseudoautosomal genes and their mouse homologs: implications for sex chromosome evolution. *Genome Res.* **11**, 2095–2100.

Graves JA, Wakefield MJ, Toder R (1998) The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Hum. Mol. Genet.* **7**, 1991–1996.

Ingman M, Kaessmann H, Paabo S, Gyllensten U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. *Trends Biochem Sci.* **23**, 403–405.

Jegalian K, Page DC (1998) A proposed pathway by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature* **394**, 776–780.

Johnson ME, Viggiano L, Bailey JA *et al.* (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519.

Joyce GF (2002) The antiquity of RNA-based evolution. *Nature* **418**, 214–221.

Kaessmann H, Zollner S, Nekrutenko A, Li W-H (2002) Signatures of domain shuffling in the human genome. *Genome Res.* **12**, 1642–1650.

Knight RD, Landweber LF (2000) The early evolution of the genetic code. *Cell* **101**, 569–572.

Koop BF, Hood L (1994) Striking sequence similarity over almost 100 kilobases of human and mouse T cell receptor DNA. *Nature Genet.* **7**, 48–53.

Krings M, Stone A, Schmitz W, Krainitzki H, Stoneking M, Paabo S (1997) Neanderthal DNA sequences and the origins of modern humans. *Cell* **90**, 19–30.

Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967.

Lahn BT, Pearson NM, Jegalian K (2001) The human Y chromosome in the light of evolution. *Nature Rev. Genet.* **2**, 207–216.

Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Hum. Mol. Genet.* **11**, 1561–1567.

Li W-H, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature* **409**, 847–849.

Logsdon Jr JM (1998) The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **8**, 637–648.

Long M (2001) Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11**, 673–680.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.

Lynch M, Richardson AO (2002) The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **12**, 701–710.

Makalowski W, Boguski MS (1998) Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl Acad. Sci. USA* **95**, 9407–9412.

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genome duplication during early chordate evolution. *Nature Genet.* **31**, 200–204.

Moran JV, DeBerardinis RJ, Kazazian HH Jr (1999) Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1504.

Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.

Muller S, Wienberg J (2001) Bar-coding primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. *Hum. Genet.* **109**, 85–94.

Needleman SB, Wunsch CD (1970) A general method applicable to the search of similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* **48**, 443–453.

O'Brien SJ, Menotti-Raymond M, Murphy WJ *et al.* (1999) The promise of comparative genomics in mammals. *Science* **286**, 479–481.

Olson MV, Varki A (2003) Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Rev. Genet.* **4**, 20–28.

Otto SP, Whitton J (2000) Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437.

Paabo S (2003) The mosaic that is our genome. *Nature* **421**, 409–412.

Patthy L (1999) Genome evolution and the evolution of exon-shuffling – a review. *Gene* **8**, 103–114.

Perry J, Palmer S, Gabriel A, Ashworth A (2001) A short pseudoautosomal region in laboratory mice. *Genome Res.* **11**, 1826–1832.

Pesce A, Bolognesi M, Bocedi A (2002) Neuroglobin and cytoglobin. *EMBO Reports* **3**, 1146–1151.

Ried K, Rao E, Schiebel K, Rappold GA (1998) Gene duplications as a recurrent theme in the evolution of the human pseudoautosomal region 1: isolation of the gene *ASMTL*. *Hum. Mol. Genet.* **7**, 1771–1778.

Rosenberg NA, Pritchard JK, Weber JL (2002) Genetic structure of human populations. *Science* **298**, 2381–2385.

Samonte RV, Eichler E (2001) Segmental duplications and the evolution of the primate genome. *Nature Rev. Genet.* **3**, 65–72.

Satta Y, Klein J, Takahata N (2000) DNA archives and our nearest relative: the trichotomy problem revisited. *Mol. Phylogen. Evol.* **14**, 259–275.

Schwartz S, Kent WJ, Smit A (2003) Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107.

Stead JDH, Jeffreys AJ (2002) Structural analysis of insulin minisatellite alleles reveals unusually large differences in diversity between Africans and non-Africans. *Am. J. Hum. Genet.* **71**, 1273–1284.

Stringer C (2002) Modern human origins: progress and prospects. *Proc. Royal Soc. Lond.* B. **357**, 563–579.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. *Proc. Natl Acad. Sci. USA* **98**, 2509–2514.

Tagle DA, Stanhope MJ, Siemieniak DR, Benson P, Goodman M, Slightom JL (1992) The β globin gene cluster of the prosimian primate *Galago crassicaudatus*: nucleotide sequence determination of the 41 kb cluster and comparative sequence analyses. *Genomics* **13**, 741–760.

Templeton AR (2002) Out of Africa again and again. *Nature* **416**, 45–51.

Thomas JW, Schueler MG, Summers TJ *et al.* (2003) Pericentromeric duplications in the laboratory mouse. *Genome Res.* **13**, 55–63.

Ureta-Vidal A, Ettwiller L, Birney E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nature Rev. Genet.* **4**, 251–262.

Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. *Nature Genet.* **27**, 422–426.

Waterman MS, Smith TF, Beyer WA (1976) Some biological sequence metrics. *Adv. Math.* **20**, 367–387.

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nature Rev. Genet.* **2**, 333–341.

Young JM, Friedman C, Williams EM, Ross JA, Tonnes-Priddy L, Trask BJ (2002) Different evolutionary processes shaped the mouse and human olfactory receptor gene families. *Hum. Mol. Genet.* **11**, 535–546.

Yunis JJ, Prakash O (1982) The origin of man: a chromosomal pictorial legacy. *Science* **215**, 1525–1530.

第13章　孟德尔性状的遗传定位

本章内容

13.1　重组体与非重组体

原则上讲，人类的遗传定位与其他任何有性繁殖二倍体有机体的遗传定位完全相同。目的是为了揭示两个基因座如何通过减数分裂重组而分离。设想一个在两个基因座为杂合性的人（基因型为 $A_1A_2B_1B_2$），假定这个人的等位基因 A_1 和 B_1 来自双亲之一，等位基因 A_2 和 B_2 来自双亲中的另一个。对这两个基因座来说，这个人的任何一个携带这些亲代组合之一的配子（A_1B_1 或 A_2B_2）是非重组体而携带 A_1B_2 或 A_2B_1 的配子是重组体（图 13.1）。重组体配子所占的比例为基因座 A 和 B 之间的重组值。

13.1.1　重组值是遗传距离的衡量标准

如果两个基因座位于不同的染色体，他们将独立地分离。考虑图 13.1 中 II_1 个体的精子发生，在减数分裂Ⅰ结束时，任何一个接受等位基因 A_1 的精子，有 50%的机会接受等位基因 B_1 而有 50%的机会接受等位基因 B_2。因此，平均来说，50%的配子将是重组体而 50%的配子将是非重组体。重组值为 0.5。如果基因座是**同线的**（syntenic），即它们位于同一染色体上，那么可能预期它们总是一起分离而没有重组体的形成。然而，这种简单的预期忽略了减数分裂过程中的交换。在减数分裂前期Ⅰ，成对的同源染色体联会，交换染色体片段（图 2.11）。只有四条染色单体中的两条涉及任一特定的交换。发生于两个基因座间的交换将产生两条携带 A_1B_2 和 A_2B_1 的重组体染色单体，并

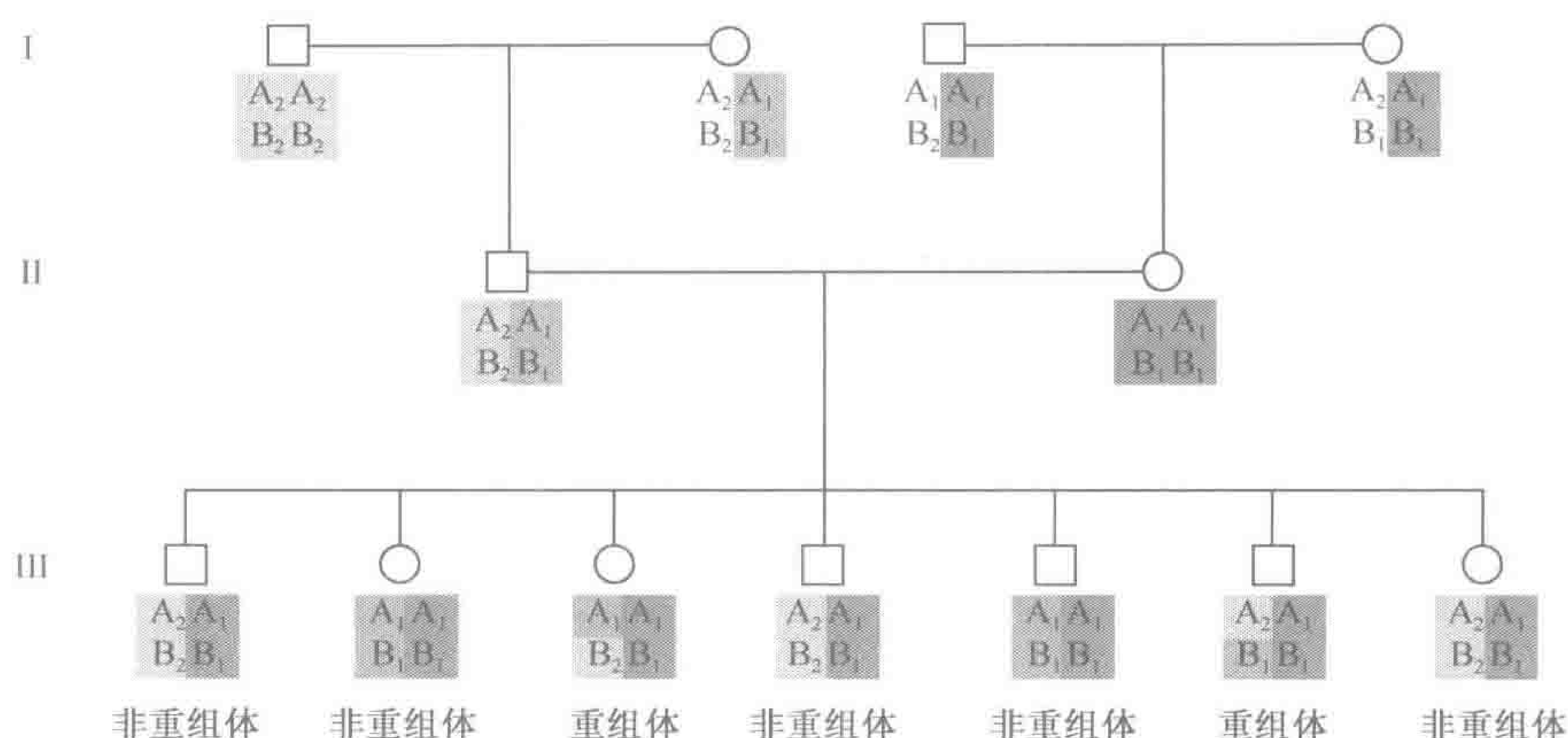


图 13.1　重组体和非重组体

在这个家系中，两个基因座的等位基因（基因座 A，等位基因 A_1 和 A_2；基因座 B，等位基因 B_1 和 B_2）是分离的。彩色框标示能够在系谱中追踪的等位基因的组合。在世代Ⅲ，我们能够区分从他们的父亲处接受非重组体精子（A_1B_1 或 A_2B_2）或者重组体精子（A_1B_2 或 A_2B_1）的人。因为Ⅱ$_2$ 个体在这两个基因座是纯合性的，所以我们无法在世代Ⅲ中鉴别哪一个体来自非重组体或重组体卵母细胞。

留下两条未涉及的非重组体染色单体。因此，一次交换在位于它两侧的基因座之间产生50%重组体。

重组很少分离一条染色体上非常紧密地在一起的基因座，因为只有精确地位于两个基因座间小空间的交换才会产生重组体。因此，在同一小染色体片段上成组的等位基因倾向于以一个板块在家系中传递。这样的等位基因板块称为**单体型**（haplotype）。单倍体标志在重组未被破坏时能够经系谱和群体追踪的、可识别的染色体片段。图 13.9 举出实例。

位于一条染色体上的两个基因座相距越远，交换将分离它们的可能性越大。因此，重组值是两个基因座间距离的衡量标准。重组值规定了**遗传距离**（genetic distance），它与**物理距离**（physical distance）不同。显示 1%重组的两个基因座在遗传图上规定为相距 1 个**厘摩**（centimorgan，cM）。

13.1.2　无论物理距离多远，重组值不超过 0.5

单个重组事件产生两条重组体染色单体和两条非重组体染色单体。当基因座很好地分离时，基因座间可能发生不止一次的交换。双重交换可涉及两条、三条或四条染色单体，但是图 13.2 显示，所有双重交换平均计算后综合的结果是产生 50%重组体。同一染色体上距离非常远的基因座可能被三次或更多的交换分离。综合的结果也是产生50%重组体。无论基因座相距多远，重组值决不超过 0.5。

13.1.3　作图函数确定重组值和遗传距离间的关系

因为重组值决不会超过 0.5，所以它们在遗传图上就不是简单地累加。如果一系列基因座，A、B、C……以 5 个 cM 的间隔分布于图上，基因座 M 可能距离基因座 A

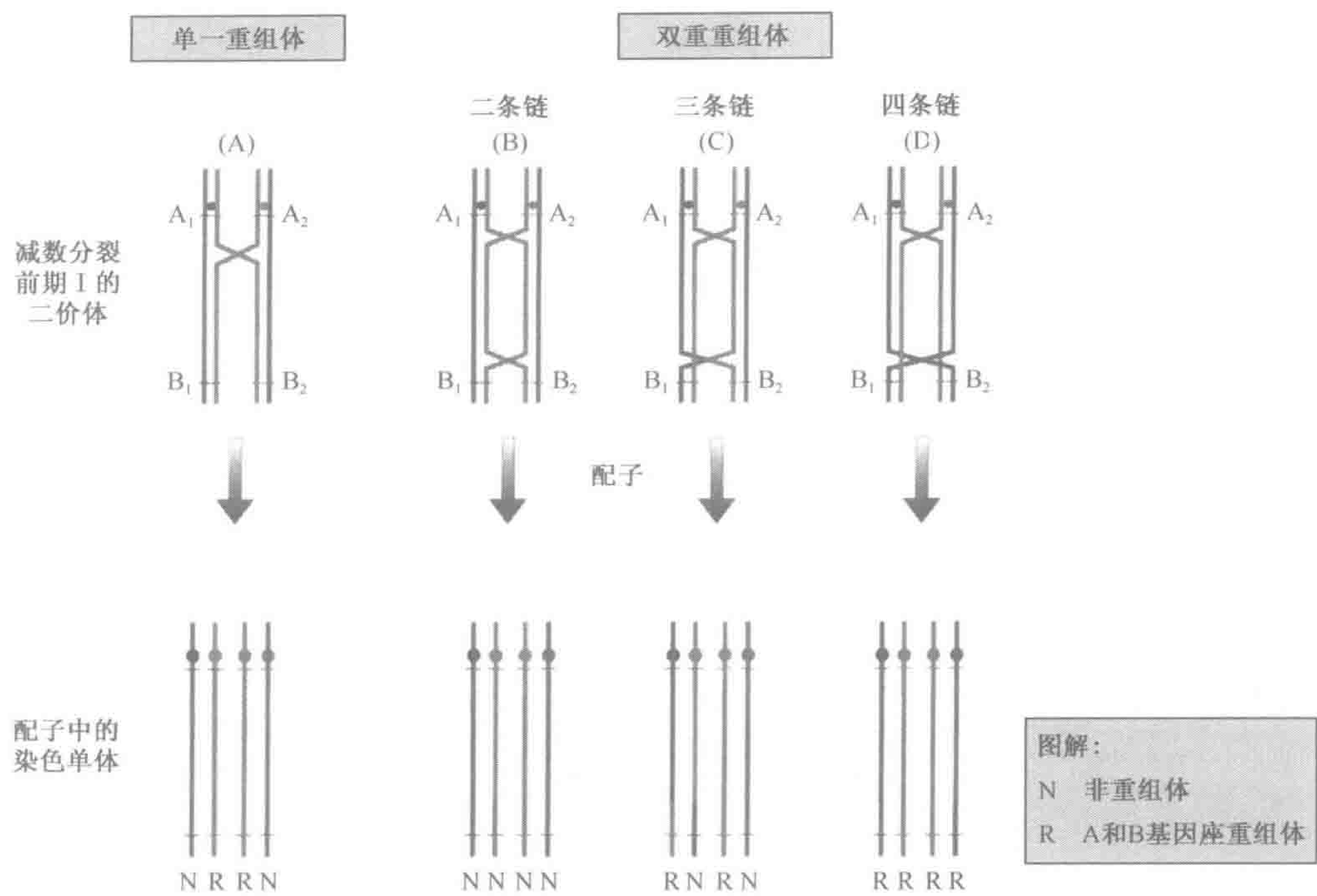


图 13.2　单一重组体和双重重组体

每次交换涉及两条联会同源染色体的四条染色单体中的两条。在两个基因座，一条染色体携带等位基因 A_1 和 B_1，另一条染色体携带等位基因 A_2 和 B_2。(A) 一次单一交换产生两条重组体染色单体和两条非重组体染色单体（50%重组体）。(B) 三种类型的双重交换以随机比例发生，所以一次双重交换的平均效应是产生 50%重组体。

60cM，但是基因座 A 与 M 之间的重组值不会是 60%。重组值和遗传图距离之间的数学关系用**作图函数**（mapping function）表示。如果沿着二价体交换随机发生且不相互影响，那么恰当的作图函数将是 Haldane's 函数：

$$w=-1/2\ln(1-2\theta)\text{或}\theta=1/2[1-\exp(-2w)]$$

其中 w 是图距而 θ 是重组值；ln 通常代表以 e 为底的对数，exp 代表以 e 为底的幂函数。然而，我们知道一个交叉的存在会抑制附近第二个交叉的形成。这种现象称为**干涉**（interference）。不同程度的干涉导致多种作图函数的存在。一个广泛用于人类作图的函数为 Kosambi's 函数：

$$w=1/4\ln[(1+2\theta)/(1-2\theta)]\text{或}\theta=1/2[\exp(4w-1)]/[\exp(4w+1)]$$

多点定位（节 13.4）需要利用作图函数将重组值的原始数据转变为遗传图。Broman 和 Weber（2000）利用大量的连锁数据评估适于人类的最佳作图函数。他们估计发生于相距 d cM 的标记间的、真实的双重交换的概率大致为 $(0.0114d-0.0154)^4$。当 d 等于 10 时，这个运算结果仅为 0.01%。感兴趣的读者可参考 Ott 的书（进一步阅读）对作图函数进行更详尽的讨论。

13.1.4　交叉计数和总的图长

减数分裂过程中的每次交换产生两条重组体染色单体和两条非重组体染色单体，在

两侧的标记间产生50%重组。因此，一次交换为整个遗传图长度贡献50cM。通过在显微镜下计数**交叉**（chiasmata）并以每个细胞中交叉的数目乘以50，我们能够以cM估计总的图长。男性减数分裂能够在睾丸活检中进行研究，每个细胞平均交叉计数为50.6（Hultén and Lindsten，1973；图13.3A），产生一个长2530cM的图。人类女性减数分裂Ⅰ发生于胎儿期16～24周，难于观察，但是最近一项新技术使交叉计数成为可能（Tease *et al.*，2002）。利用荧光抗体标记一种蛋白——MLH1的位置，而MLH1是重组结构的一部分（图13.3B）。交叉更常见于女性减数分裂（证明了关于异形配子的性别具有较低交叉计数的Haldane规则）。在一个流产的女性胎儿的卵巢中平均的常染色体交叉计数为70.3，产生一个长3515cM的图。这些细胞学衍生的图长能与来自遗传作图得到的2590cM（男性）和4281cM（女性）的最佳估算相比（Kong *et al.*，2002）。唯独在假常染色体区域之外（节2.3.3）的Y染色体没有遗传图，因为在正常的减数分裂过程中Y染色体不发生联会与交换。

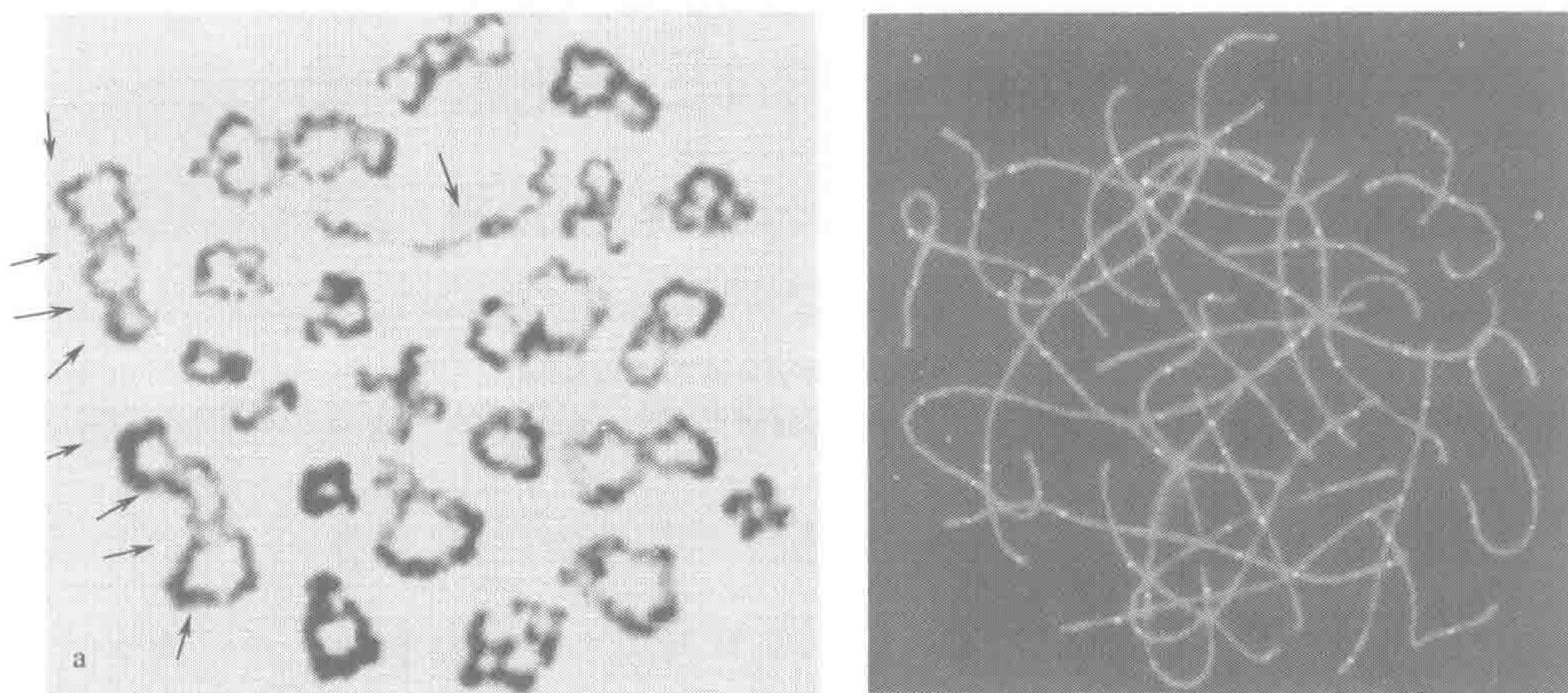


图 13.3 男性和女性减数分裂的交换

（A）男性减数分裂：中期Ⅰ的精母细胞。注：X和Y染色体的端对端配对。交叉标示每个二价体内交换的位置（箭头）。（B）女性减数分裂粗线期。MLH1荧光抗体的亮点标示交换的位置。伯明翰Maj Hultén教授提供照片。经Nature Publishing Group允许出自Hultén和Tease（2003）。

13.1.5 物理图与遗传图：重组体的分布

物理图表示沿染色体排列的特征顺序并以kb或Mb描述其距离；遗传图表示他们的顺序及其因重组而分离的概率。既然两张图上的特征顺序应该相同，若在所有的染色体位置每Mb DNA重组的概率一致，那么距离将只能相符。实际上，重组概率根据性别和染色体位置的不同而有相当大的变化。个体每次减数分裂交换的平均数不同，如确实这样，那么他们也会有不同的遗传图长度。

既然我们有人类基因组序列，那么微卫星或SNP标记就可以通过寻找使用的PCR引物的匹配序列而在序列数据库中进行物理图定位。因此，沿着一条染色体的重组的分布可直接在显微镜下估计（图13.3），或者通过标记的遗传图距离与物理距离的关系而估计。两种方法均显示重组不是随机的。男性染色体的端粒位置更多见重组，女性着丝

粒区有重组体而男性没有（图 13.4 与 Kong *et al*.，2002）。如上所述，干涉影响双重重组体的间隔。作为一个大致粗略的估计，1cM＝1Mb，但是存在长达 5Mb 而性别-平均的重组每 Mb 小于 0.3cM 的重组“沙漠”，以及每 Mb 大于 3cM 的重组“丛林”。X 和 Y 染色体短臂末端的假常染色体区域显示了最极端的偏差（节 12.2.7）。在此 2.6Mb 区域内，男性具有一专性的交换，故长为 50cM。因此，在此区域男性 1Mb＝19cM，而女性 1Mb＝2.7cM。

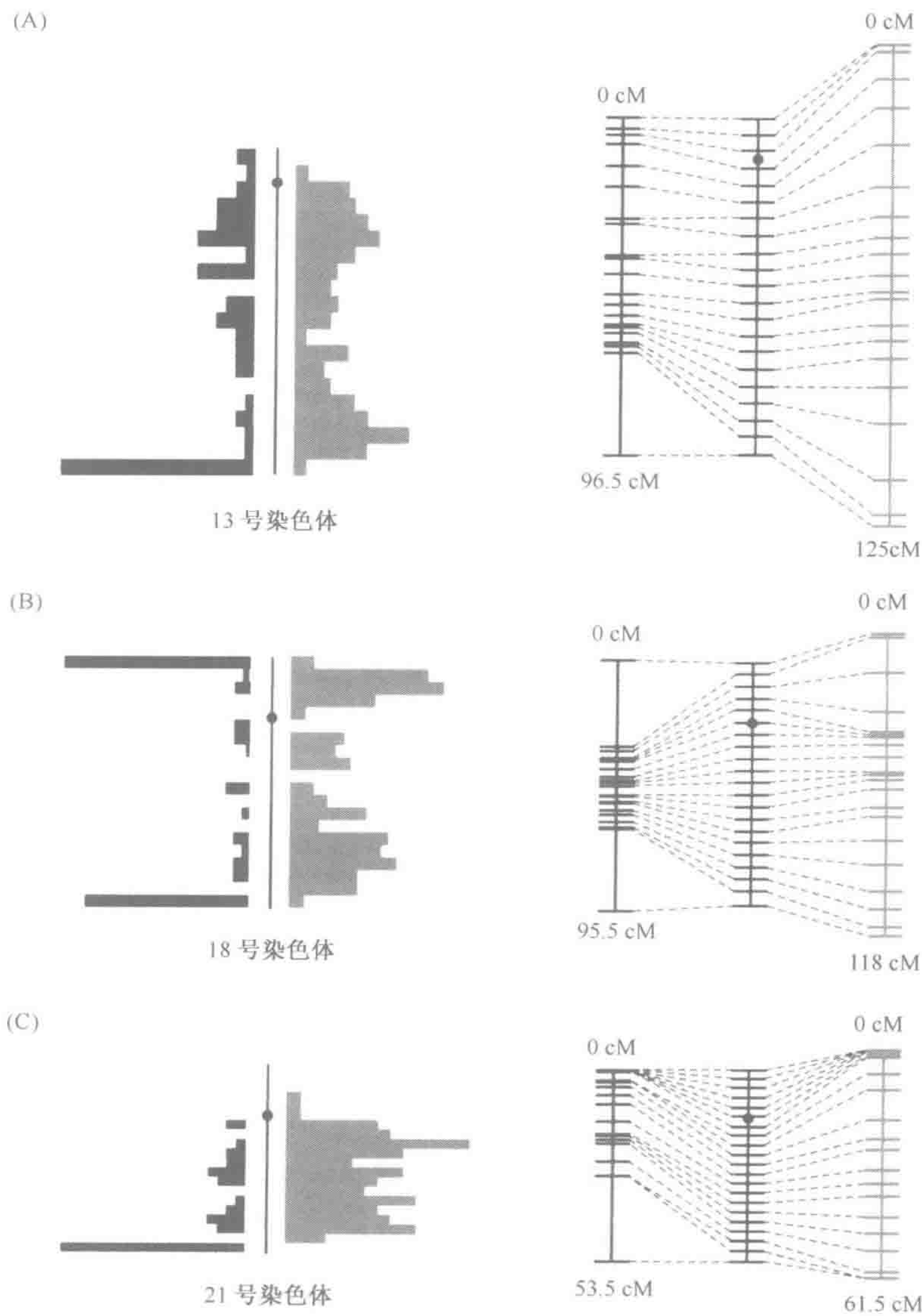


图 13.4　重组体的分布是非随机的、性别特异的

直方图说明了 13 号、18 号和 21 号染色体上精母细胞中交叉的分布（黑色）和卵母细胞中 MLH1 焦点的分布（浅灰色）。每个染色体对以其长度的 5%为间隔划分。每条染色体上这些交换的分布模式也以重组图的形式表现出来。这些图强调了男性和女性生殖细胞中交换数目与分布的不同模式及在重组图上的必然结果。

经 Nature Publishing Group 允许出自 Hultén 和 Tease（2003）。

最近的研究也表明在 DNA 序列水平上重组是非随机的。如在节 15.4.3 所描述的，我们的染色体看起来好像是由保守的板块组成，典型的板块为 20～50kb 长，被重组热点分隔。大约全部重组的 95%发生在这些 1～2kb 的热点内（图 13.5；Jefferys *et al.*，2001）。而且，Jeffreys 对少数几个这样的热点的详细描述提示重组总是开始于同一点的 10bp 内。令人着急的是，这些点的 DNA 序列没有任何明显的专有特征可以解释这种现象。

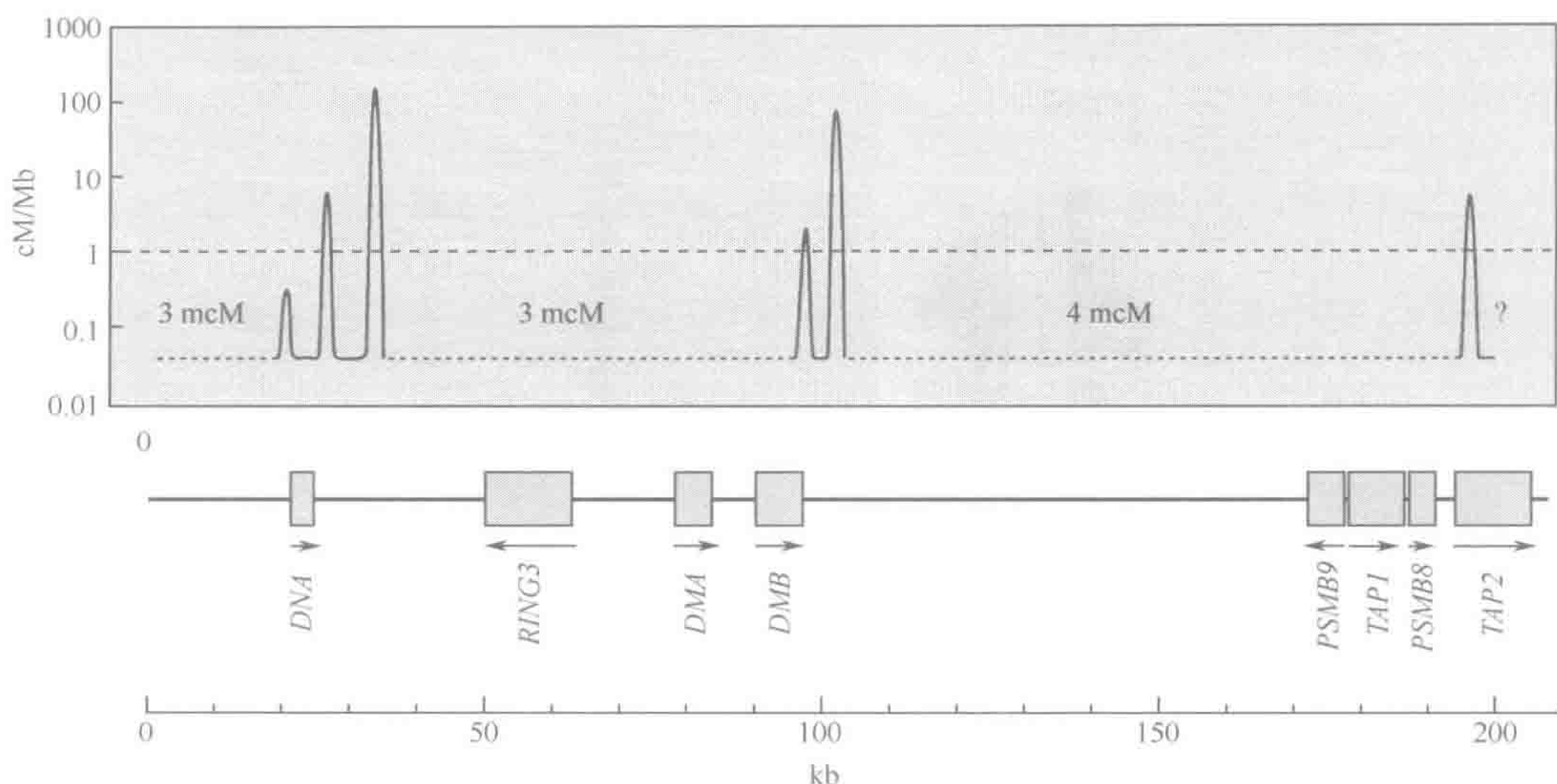


图 13.5 重组集中于少数热点

在男性的主要组织相容性复合体区域 DNA 水平重组的分布。全部重组的 95%均发生在高度集中的热点区。经 Nature Publishing Group 允许再引自 Jeffreys 等（2001），Nat. Genet. 29，217～222。

13.2 遗传标记

13.2.1 定位人类致病基因需要遗传标记

因为大多数人类遗传学家对疾病感兴趣，所以我们便会喜欢一张显示所有致病基因的顺序及彼此距离的图。计算疾病基因对间的重组值将会是构建这张图的显而易见的途径，但是在人类，疾病-疾病定位是不可能的。如我们所见（图 13.1），遗传定位需要双重杂合子。表现出两种不同疾病杂合性的人非常罕见。即使能够找到他们，他们也可能没有孩子，或者因为其他一些原因不适于遗传分析。正因为如此，人类遗传定位取决于**标记**（marker）。标记是指任何具有多态性的、能够在系谱中用于追随染色体片段的孟德尔性状。如果能够利用容易获得的材料（血细胞而不是脑活检）简便地、廉价地区分标记，这个标记是有帮助的，但重要的是它应该是高度多态的，随机选择的个体（在此标记处）有很大的机会是杂合性的。框 13.1 总结了人类遗传标记的发展，从血型至现在的 DNA 微卫星和单核苷酸多态性。

框 13.1　人类遗传标记的发展

标记类型	基因座数目	特　征
血型 1910～1960	～20	需要新鲜血，稀有的抗血清 由于显性，总是无法从表型推断基因型 不容易物理定位
血清蛋白电泳迁移率变异体 1960～1975	～30	需要新鲜的血清，专门的检测 不容易物理定位 常常有限的多态性
HLA 组织类型 1970～	1（单体型）	一个连锁组 高度提供信息的 只能用于检测与 6p21.3 的连锁
DNA RFLP（图 7.5A，图 7.6） 1975～	$>10^5$（潜在地）	两个等位基因标记，最大杂合度 0.5 最初需要 Southern 印迹杂交，现在为 PCR 容易物理定位
DNA VNTR（小卫星）（图 7.5B） 1985～	$>10^4$（潜在地）	许多等位基因，高度提供信息的 通过 Southern 印迹杂交分型 容易物理定位 趋向成簇分布于染色体末端
DNA VNTR（微卫星）（图 7.7） （二、三和四核苷酸重复） 1989～	$>10^5$（潜在地）	许多等位基因，高度提供信息的 可以用全自动多重 PCR 分型 容易物理定位 分布于整个基因组
DNA SNPs 单核苷酸多态性	$>4\times10^6$	信息含量少于微卫星 可以大规模地通过自动化仪器分型而不需要凝胶电泳

VNTR，可变数目串联重复

疾病-标记定位，如果不是一个纯粹的盲目试验，需要标记的框架图。框架图由标记-标记定位产生。虽然理论上相距 40cM 的基因座间可以检测到连锁，但所需要的数据量是受到限制的。如果没有重组体，10 次减数分裂足以证明连锁，但如果重组值为 0.3，则需要 85 次减数分裂才能提供同样强的连锁证据（这些计算的指南见框 13.3）。对于一种罕见的疾病，获得足够的家系材料并检测多于 20～30 次减数分裂是很困难的。因此，定位需要基因组中彼此间隔不大于 10～20cM 的标记。假定上述计算的基因组长度，考虑到不完全的信息量（见下文），我们最少需要几百个标记。人类基因组计划早期的重大成就是产生了 10 000 个以上高度多态的微卫星标记并把它们置于框架图中（Broman *et al.*，1998）。这些图使 20 世纪 90 年代定位孟德尔疾病的惊人进展成为可能。比较新的疾病关联研究（节 15.4）需要密度更高的标记图，这种需求随着几百万个 SNP 标记的发展而得以解决。

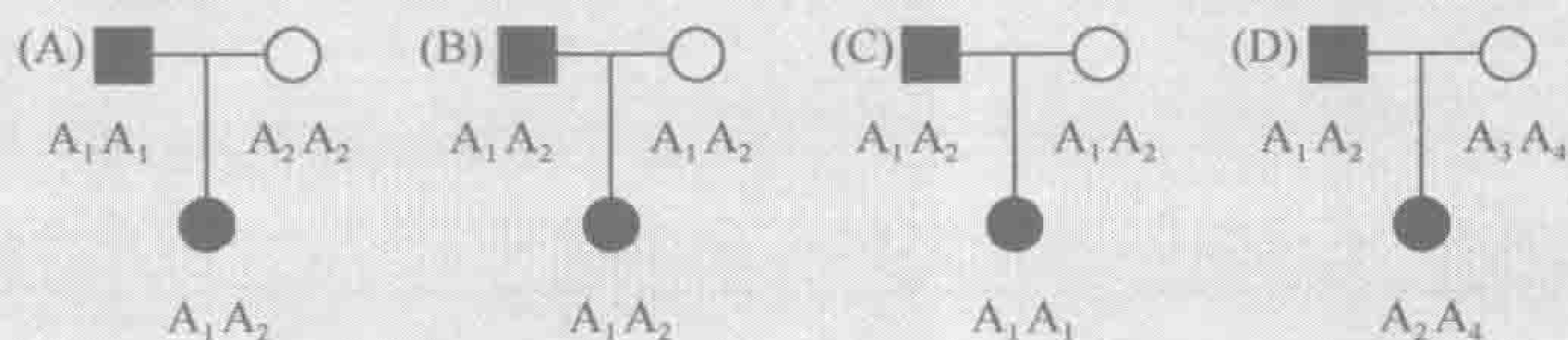
13.2.2　杂合度或多态信息含量衡量一个标记是如何提供信息的

对于连锁分析，我们需要**提供信息的减数分裂**（informative meiosis）（框 13.2）。

框中的例子显示若双亲在一既定的标记处是纯合性的，那么对于该标记（来说），减数分裂是不提供信息的；在双亲具有同一杂合基因型的病例中有一半也是不提供信息的。对于大多数用途，一个标记的平均**杂合度**（heterozygosity）（随机选择的个体是杂合性的几率）是用来作为信息的衡量标准的。如果标记等位基因 A_1、A_2、A_3……，各自的基因频率为 p_1、p_2、p_3……，那么杂合性个体的比例为 $1-(p_1^2+p_2^2+p_3^2+\cdots\cdots)$（节 4.5.1）。一个更复杂的但极少使用的衡量标准，**多态信息含量**（polymorphism information content，PIC），考虑到病例中某些人是杂合性的但不提供信息的情况，就像框 13.2 中系谱 B。

框 13.2 提供信息的和不提供信息的减数分裂

当我们能够鉴定配子是否为重组体时，一个减数分裂对于连锁是提供信息的。考虑下面四个系谱中对孩子产生父方参与的男性减数分裂。父亲具有与标记等位基因 A_1 一起遗传的显性疾病。他将这种疾病传递给他的女儿——但我们能够判断精子是否为疾病基因与标记基因座之间的重组体吗？



（A）这个减数分裂是不提供信息的：无法辨别纯合性父亲的标记等位基因。（B）这个减数分裂是不提供信息的：孩子可能从父亲遗传 A_1，从母亲遗传 A_2，或者相反。（C）这个减数分裂是提供信息的且没有重组体：孩子从父亲遗传 A_1。（D）这个减数分裂是提供信息的且有重组体：孩子从父亲遗传 A_2。

13.2.3 DNA 多态性是现在所有遗传标记的基础

在 20 世纪 80 年代早期，DNA **多态性**（DNA polymorphism）第一次提供了一套数目众多、广泛分布于整个基因组并可用于全基因组寻找连锁的标记。DNA 标记另有一个优势，即可以全部采用相同的技术对他们进行分型。而且，通过辐射杂种细胞作图（框 8.4）或在人类基因组序列中寻找与 PCR 引物匹配获得的标记以确定它们的染色体位置。这使得以 DNA 为基础的遗传图与物理图相互参考，避免长期寻找的囊性纤维化基因（*CFTR*）初次定位时的挫折。与二乙基对硝基苯磷酸酯酶的蛋白质多态性建立了连锁，但二乙基对硝基苯磷酸酯酶基因的染色体位置却是未知的。DNA 标记的发展使人类基因定位实实在在地开始了。

限制性片段长度多态性（RFLP）

第一代 DNA 标记是**限制性片段长度多态性**（restriction fragment length polymorphism，RFLP）。RFLP 最初是通过待测 DNA 经限制酶消化并与放射标记探针进行 Southern 印迹杂交来分型的（图 7.5）。这项技术需要大量的时间、金钱和 DNA，使得全基因组搜索成为英雄的事业。现在这不再是问题，因为 RFLP 通常能用 PCR 来分型。

一段含有可变的限制酶切位点的序列经过扩增后，产物与适当的限制酶共同孵育，然后经凝胶电泳观察它是否被切开（图 7.6）。一个更主要的不足之处是它们的低信息含量。RFLP 只有两个等位基因：（限制酶切）位点存在或不存在。最大杂合度为 0.5。利用 RFLP 进行疾病定位常因在一个家系中关键的减数分裂不能提供信息而失败。

小卫星

小卫星 VNTR（可变数目串联重复）标记是一个很大的改进。VNTR（节 7.1.3）有许多等位基因，且杂合度高。大多数减数分裂是提供信息的。但是，Southern 印迹杂交和放射性探针的技术问题仍是容易定位的阻碍，并且 VNTR 不是均匀地分布于基因组。

微卫星

PCR 的出现最终使定位相对迅速和容易。小卫星太长以至于不能很好地扩增，因此 PCR 连锁分析的标准工具是**微卫星**（microsatellite）。大部分是$(CA)_n$重复。三核苷酸重复和四核苷酸重复逐渐代替了二核苷酸重复而成为选择的标记，因为它们能够给出更清晰的结果。二核苷酸重复序列在 PCR 扩增过程中特别容易产生复制滑动（节 11.3.1），以致每个等位基因在凝胶上产生一个小的**“梯状伪带”**（ladder of stutter band），难以阅读（图 7.8）。更多的努力致力于产生可相互兼容的成套微卫星标记，它们能够在多重 PCR 反应中一起扩增，产生不重叠的等位基因片段长度，因此可在同一泳道内进行电泳。通过标记不同颜色的荧光，就有可能在一个全自动凝胶电泳的单一泳道内进行一个样本大约 10 个标记的检测。

单核苷酸多态性（SNP）

经过发展越来越多的多态标记的 10 年之后，看起来反常的是最新一代标记是二等位基因的**单核苷酸多态性**（single nucleotide polymorphism，SNP）。它们包括了经典的 RFLP，也包括了那些并非碰巧产生或取消限制酶切位点的多态性。SNP 的优势在于超高通量的基因分型和极高的标记密度（Wang *et al.*，1998）。寻找疾病易感基因（第 15 章）需要利用间隔非常近的标记分析大量的基因型。基因组中的微卫星不足以在每 10kb 或更少距离就有一个标记提供所需密度。而且，微卫星是通过凝胶电泳分析，这就限制了通量。一个国际性的学术界与工业界合作的协作组已经产生了一个公共的、拥有超过 400 万 SNP 的数据库（dbSNP，可通过 NCBI 网站进入），并发展了超大规模的、无需凝胶电泳即可进行 SNP 分析的方法（节 18.4.2）。

13.3 两点定位

13.3.1 在人类系谱中计数重组体通常不是简单的

收集了孟德尔疾病家系并利用提供信息的标记进行分型后，我们如何知道何时已经发现了连锁？这个问题有两个方面：

1. 我们如何计算出重组值？
2. 我们应该使用何种统计学检验判定重组值与没有连锁存在的无效假设的预期值

0.5 之间是否存在显著性差异？

在一些家系中，第一个问题可通过计数重组体与非重组体而很简单地回答。图 13.1 中的家系就是一个例子。在 7 次减数分裂中有 2 个重组体，重组值为 0.28。图 13.6A 表示了另一个例子。能够为连锁提供信息的双重杂合子（图 13.1 和图 13.6A 中的 $Ⅱ_1$ 个体）是状态已知的：我们知道哪个等位基因遗传自哪位亲本，因此，我们能够清楚地判断每次减数分裂是重组体还是非重组体。在图 13.6B，$Ⅱ_1$ 个体也是双重杂合性的，但这次是状态未知的。在她的孩子中，要么有五个非重组体和一个重组体，要么有五个重组体和一个非重组体。即使第一种情况看起来比第二种情况更有可能，我们也不能清楚地鉴定出重组体。图 13.6C 增加了更多的复杂性——但是如果这是一个具有罕见疾病的家系，没有任何一个研究人员愿意放弃它。需要一些方法从这样不完整的家系中提取连锁信息。

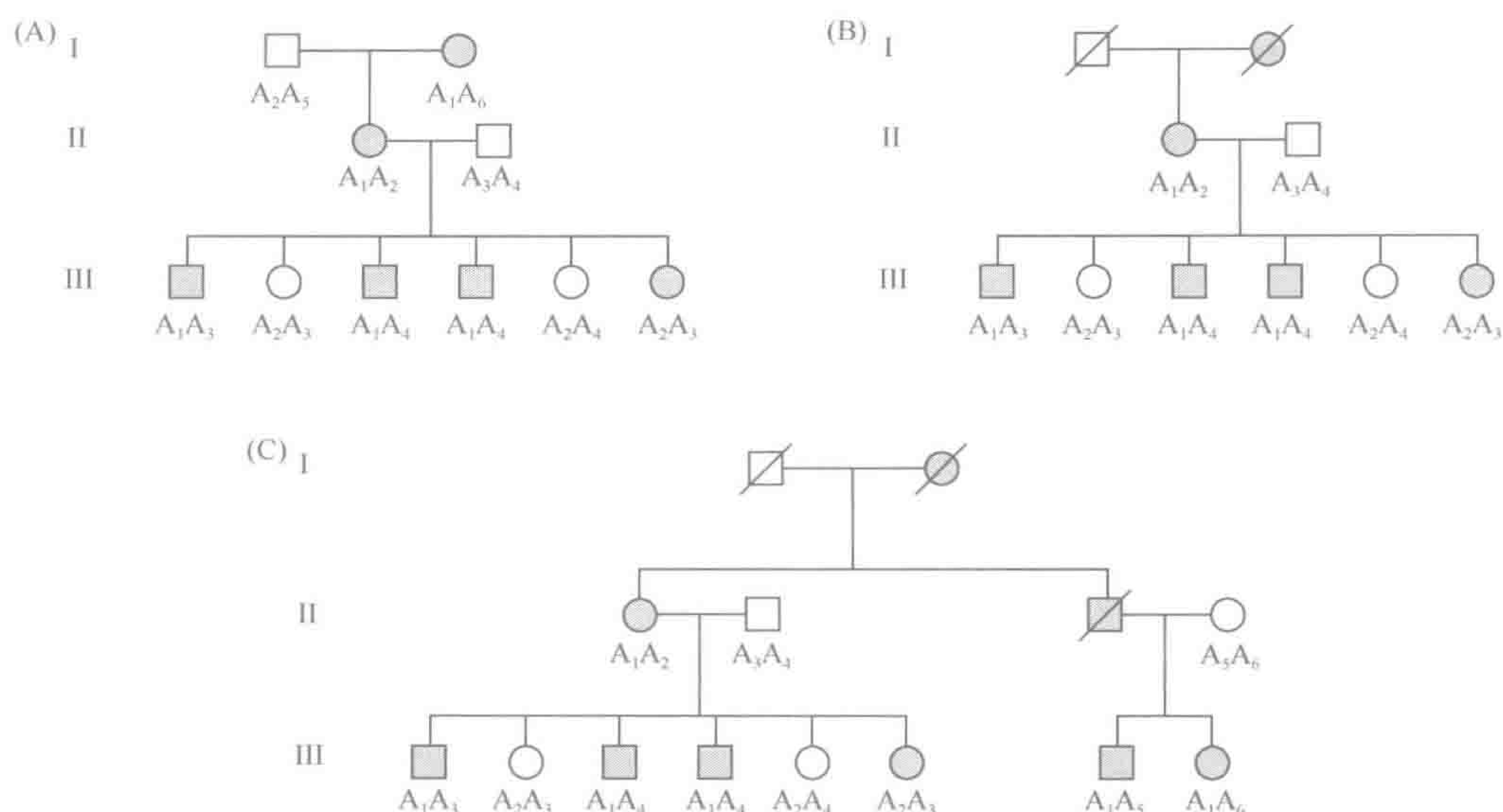


图 13.6 识别重组体

一个常染色体显性遗传病家系的三个版本，以标记 A 表示。(A) 所有的减数分裂状态是已知的。我们能够明白地鉴定 $Ⅲ_1 \sim Ⅲ_5$ 为非重组体，而 $Ⅲ_6$ 为重组体。(B) 同一个家系，但状态却是未知的。母亲 $Ⅱ_1$ 可能遗传伴随疾病的标记等位基因 A_1 或 A_2，因此，她的状态是未知的。要么 $Ⅲ_1 \sim Ⅲ_5$ 为非重组体，而 $Ⅲ_6$ 为重组体；要么 $Ⅲ_1 \sim Ⅲ_5$ 为重组体，而 $Ⅲ_6$ 为非重组体。(C) 经进一步追踪亲属后的同一个家系。$Ⅲ_7$ 和 $Ⅲ_8$ 也从他们的父亲遗传了伴随疾病的标记等位基因 A_1，但是我们无法确定他们父亲的等位基因 A_1 是否与其父的姐姐 $Ⅱ_1$ 的等位基因 A_1 具有传递一致性。也许在四个祖父辈标记等位基因中存在等位基因 A_1 的两个拷贝。这种似然性主要取决于等位基因 A_1 的基因频率。因此，尽管这个家系中含有连锁信息，但是提取该信息仍存在问题。

13.3.2 计算机化对数优势比分析是分析复杂系谱孟德尔性状间连锁的最佳方法

在图 13.6B 的家系中，清楚地鉴定重组体并计数他们是不可能的。然而，假定基因座是连锁的（重组值$=\theta$）或不连锁的（重组值$=0.5$），我们就有可能计算出此家系总体的似然性。这两个似然性的比率产生连锁的优势，此优势的对数就是**对数优势比**(lod score)。Morton（1995）证明对数优势比代表了评价家系连锁的最有效的统计学方

法，并推导出用于各种标准系谱结构计算对数优势比（以 θ 的函数表示）的公式。框13.3表示了对于一个简单的结构是如何计算的。作为重组值的函数，对数优势比是在一定 θ 值范围内计算的。最可能的重组值是指在对数优势比最大处的值。在一组家系中，总的连锁概率是每一单个家系的概率的产物，因此对数优势比（以对数状态）能够在家系间累加。

框 13.3 图 13.6 家系对数优势比的计算

- 假定基因座是真正连锁的，重组值为 θ，是重组体的减数分裂的似然性为 θ，而非重组体的减数分裂的似然性为 $1-\theta$。
- 如果基因座实际上是不连锁的，是重组体或是非重组体的减数分裂的似然性为1/2。

家系 A：

有五个非重组体和一个重组体。

假定连锁，总的似然性为 $(1-\theta)^5 \cdot \theta$。

假定没有连锁，似然性为 $(1/2)^6$。

似然性比率为 $(1-\theta)^5 \cdot \theta/ (1/2)^6$。

对数优势比Z是似然性比率的对数。

θ	0	0.1	0.2	0.3	0.4	0.5
Z	$-\infty$	0.577	0.623	0.509	0.299	0

家系 B：

II_1 状态是未知的。

如果她伴随疾病遗传 A_1，就有五个非重组体和一个重组体。

如果她伴随疾病遗传 A_2，就有五个重组体和一个非重组体。

总的似然性为 $1/2[(1-\theta)^5 \cdot \theta/(1/2)^6]+1/2[(1-\theta) \cdot \theta^5/(1/2)^6]$。这考虑到任一可能的状态，具有相等的前概率。

对数优势比Z是似然性比率的对数。

θ	0	0.1	0.2	0.3	0.4	0.5
Z	$-\infty^*$	0.276	0.323	0.222	0.076	0

家系 C：

在这一点上，非自虐者求助于计算机。

计算图13.6C家系完整的对数优势比比较困难。为了计算 III_7 和 III_8 是重组体或非重组体的似然性，我们必须获得由 I_1、I_2 和 I_3 每一种可能的基因型计算出的似然性，用那种基因型的概率加权。对于 I_1 和 I_2 来说，基因型概率依赖于基因频率与观察到的 II_1、III_7 和 III_8 的基因型。然后可依据简单的孟德尔规则计算 I_3 的基因型概率。除了非常简单的病例，只要给予系谱数据和基因频率表，人类连锁分析完全依赖于履行算法的计算机程序来处理这些基因型概率的分支树。

13.3.3 对数优势比为+3和−2是连锁和排除连锁的标准（对单一的检验）

连锁分析的结果是一个不同的重组值所对应的对数优势比表，就像框13.3的两个表。正对数优势比提供支持连锁的证据而负对数优势比提供反对连锁的证据。注意只有

重组值介于 0 与 0.5 之间才有意义，所有的对数优势比当 $\theta=0.5$ 时均为 0［因为它们衡量两个一致性概率的比，并且 $\log_{10}(1)=0$］。结果可以绘制成如图 13.7 的曲线。

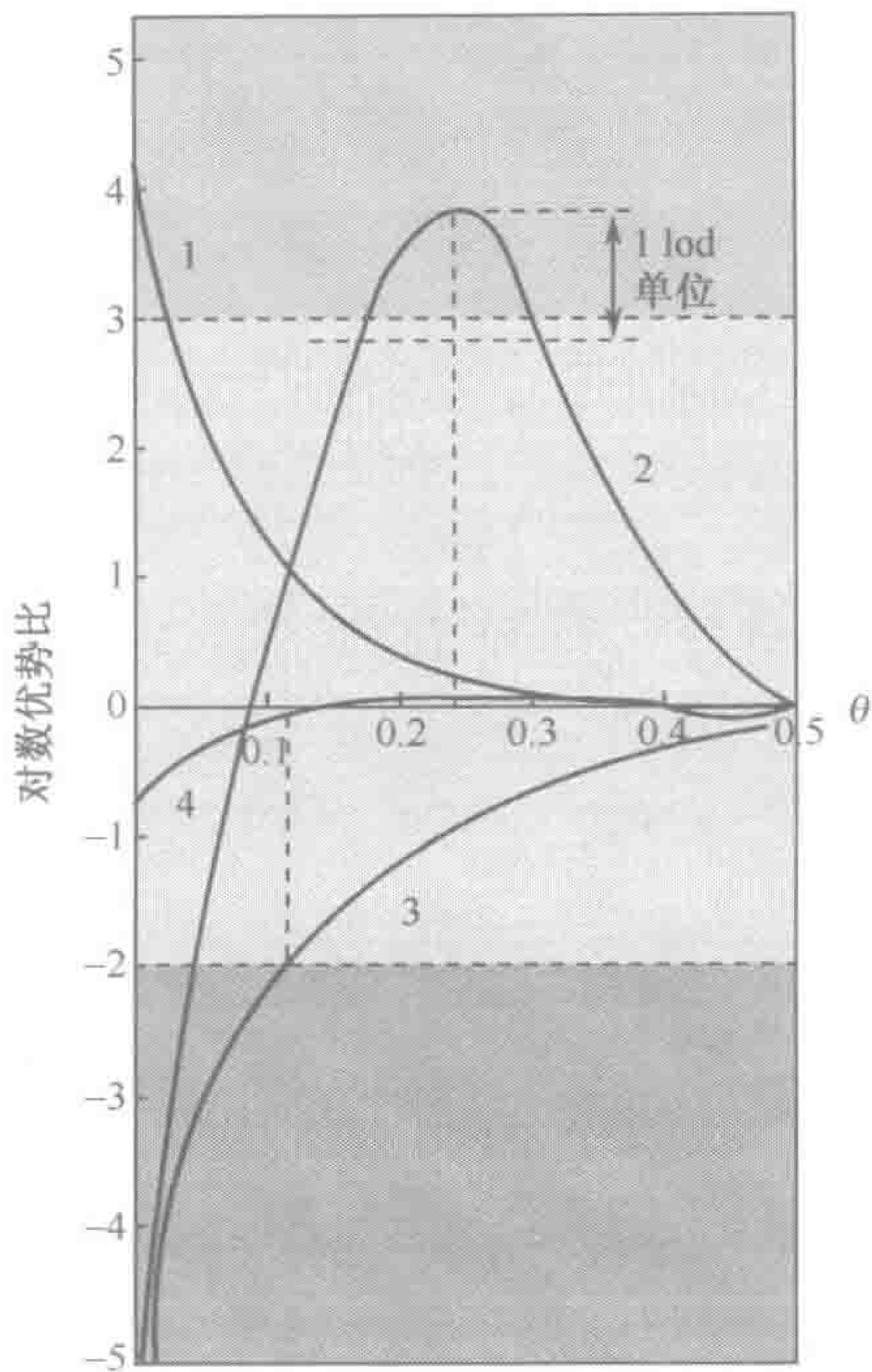


图 13.7　对数优势比曲线

由一套假定的连锁实验得到的以重组值为对照的对数优势比曲线图。曲线 1，无重组体时存在连锁（$Z>3$）。曲线 2，当最可能的重组值为 0.23 时存在连锁（$Z>3$）。曲线 3，重组值小于 0.12 时排除连锁（$Z<-2$）；重组值大于 0.12 时无法确定存在连锁。曲线 4，在所有的重组值处无法确定存在连锁。

回到本节开始时提出的两个问题，我们现在明白了最可能的重组值是对数优势比最高处的重组值。如果没有重组体，对数优势比在 $\theta=0$ 时最大。如果有重组体，Z 将在最有可能的重组值（图 13.6A 家系为 0.167＝1/6；但是图 13.6B 家系较难预测）处达到顶点。

第二个问题涉及显著性的阈值。在这里，第一眼看去答案令人吃惊，$Z=3.0$ 是接受连锁的阈值，误差几率为 5%。如果 $Z<-2.0$，可以排除连锁。Z 值介于 −2 和 +3 之间是没有确定结果的。对于大多数统计学来说，$p<0.05$ 常用作显著性的阈值，但是 $Z=3.0$ 相当于 1000：1 的优势［$\log_{10}(1000)=3.0$］。选择如此严格的阈值的原因在于随机选择的两个基因座必然连锁的内在不确定性。由于有 22 对常染色体可供选择，它们不可能位于同一染色体上（同线的），即使它们位于同一染色体上，很好分离的基因座间是不连锁的。通常的感觉告诉我们，如果一些事情具有内在的不确定性，我们需要强烈的证据说服我们它是正确的。这种通常的感觉在 Bayesian 计算（框 13.4）中能够得到定量化，Bayesian 计算显示 1000：1 优势实际上精确地相当于常规的 $p=0.05$ 的显著性阈值。同一推理表明 Lod 阈值为 2.3 是建立 X 连锁性状与 X 染色体标记之间连锁的阈值（连锁的前概率≌1/10）。

框 13.4　连锁阈值的 Bayesian 计算法

两个基因座必然连锁的似然性（连锁的前概率）曾被普遍争论，但大约 1/50 的估计值被广泛接受。

假设	基因座是连锁的（重组值＝θ）	基因座是不连锁的（重组值＝0.5）
前概率	1/50	49/50
条件概率：连锁的 1000：1 优势［Lod 值 $Z(\theta)=3.0$］	1000	1
联合概率（前概率×条件概率）	20	～1

由于两个随机选择的基因座应该连锁的前概率低，所以为了提供全部支持连锁的 20：1 优势，就需要提供支持连锁的 1000：1 优势的证据。这相当于常规的、具有统计学显著性的 $p=0.05$ 阈值。此计算是一个利用 Bayes' 公式组合概率的例子（框 18.4 和图 18.15）。对数优势比的描述见正文。

可信区间是难以分析性地推导出来，但是一个广泛接受的支持区间扩展至对数优势比低于峰值一个单位（lod-1 规则）处的重组值。因此，图 13.7 曲线 2 在重组值最可能为 0.23、支持区间为 0.17～0.23 时提供了支持连锁的证据（$Z>3$）。数据量越大，曲线越急剧地达到高峰，但是一般峰是相当宽阔的。重要的是记住人类遗传图上的距离通常是很不精确的估计值。

负对数优势比在 $Z<-2$ 的区域排除连锁。图 13.7 曲线 3 在标记两侧 12cM 内排除疾病。尽管基因定位者希望一个正对数优势比，但是排除不是没有价值的。它们告诉我们哪里没有疾病（**排除定位**，exclusion mapping）。这能够排除一个可能的候选基因，如果足够的基因组被排除了，那么仅能留下几个可能的位置。

13.3.4　一个基因组范围的显著性阈值必须用于全基因组扫描

在疾病研究中，一些家系一个接一个标记地分型，直到获得正对数优势比。恰如其分的显著性阈值是在全基因组扫描过程中，任何位置出现假阳性结果的几率只有 0.05 时的对数优势比。如框 13.4 所示，3.0 的对数优势比相当于在一单个点 0.05 的显著性。但是如果使用了 50 个标记，假阳性结果几率大于只使用一个标记时的几率。严格的过程（Bonferroni 校正）应该是在检验它的显著性之前将 p 值乘以 50。使用 n 个标记的研究，对数优势比的阈值应该是 $3+\log(n)$，即使用 10 个标记时对数优势比为 4，使用 100 个标记时对数优势比为 5，等等。然而，这是过于严格的。连锁数据不是独立的：如果一个位置被排除，那么此性状定位于另一个位置的前概率则增加。对 0.05 的基因组-广泛的显著性阈值水平已有过多的争论，但对于孟德尔性状，一个广为接受的答案是 3.3（Lander and Schork，1994）。对于非孟德尔性状，见节 15.3.4。实际上，无论使用一个标记或者许多标记，对数优势比低于 5 都应该认为是暂时的。

13.4　多点定位比两点定位更有效

13.4.1　多点连锁能够定位疾病基因座至标记的框架上

如果同时分析两个以上基因座的数据，连锁分析将更有效。多点分析对建立一组连锁的基因座的染色体顺序特别有用。为此目的，实验遗传学家长期使用三点杂交。最罕见的重组体类型是需要双重重组的一类。在表 13.1，基因顺序 A-C-B 是显而易见的。这一过程比一系列两点杂交中独立地估计 A-B、A-C 和 B-C 间隔的重组值更有效。理想地，在任何连锁分析中，应该筛查整个基因组寻找连锁，而且全部的数据都用于计算在基因组中每个位置的似然性。

在人类多基因座定位的第二个优势在于它有助于克服因标记有限的信息含量所造成的问题。在一个家系中，有些减数分裂可能在标记 A 处是提供信息的，而其他减数分裂在标记 A 处是不提供信息的但在附近的标记 B 处是提供信息的。只有同时在标记 A 与 B 处进行疾病的连锁分析才能提取全部的信息。选择高度提供信息的微卫星标记而不是二等位基因的 RFLP 用于定位这一点并不重要，但是当使用 SNP 时就不一样了。

表 13.1　通过三点杂交进行基因排序

后代种类	重组位置(×)	数目	后代种类	重组位置(×)	数目
ABC/abc	非重组体	853	Abc/abc	A-×-(B,c)	47
abc/abc			aBC/abc		
aBc/abc	(A, B) -×-C	5	AbC/abc	B-×-(A,C)	95
abC/abc			aBc/abc		

在三个连锁基因座为杂合性的小鼠(*ABC/abc*)与三重纯合子(*abc/abc*)小鼠间建立杂交。后代最罕见的类型将是那些产物需要两次交换的种类。在 1000 只动物中，142(95＋47)只为 A 与 B 之间的重组体，52(47＋5)只为 A 与 C 之间的重组体，100(95＋5)只为 B 与 C 之间的重组体。只有 5 只动物是 A 与 C 之间而不是 B 与 C 之间的重组体，所以这些动物必然具有双重交换 A-×-C-×-B。因此，图的顺序为 A-C-B，并且遗传距离约为 A-(5cM)-C-(10cM)-B。

13.4.2　标记框架图：CEPH 家系

多点定位排列基因座的能力特别有利于构建标记框架图。然而，在这样的图上排列基因座并非易事。对于 n 个标记来说，有 $n!/2$种可能的排列顺序，而现在的图上每条染色体有数以百计的标记。在人类基因组序列完成之前，这对于作图者是非常困难的。某些比无理性的计算方法更聪明的方法不得不用于计算正确的排列顺序。即使没有大规模的序列数据，物理定位信息也是非常有益的。能够通过 PCR 进行分析的标记可用作序列标签位点（STS：框 15.4），并可通过数据库搜索或者利用实验性辐射杂种细胞（框 8.4）方法进行物理定位。结果是一个物理上锚定的标记框架。

疾病-标记定位面临利用什么样的家系才能够发现感兴趣的疾病正在分离的必要性。这样的家系很少具有理想的结构。所有家系减数分裂的数目经常是少得太不合需要，且一些家系是状态未知的。标记-标记定位能够避免这些问题。标记可在任一家系中进行研究，因此，可选择有许多孩子和适于连锁分析的理想结构的家系，如图 13.1 的家系。标记框架图的构建极大地受益于巴黎人类多态研究中心家系（现在是 Jean Dausset 基金会）为此目的特别集合的家系标本（**CEPH 家系**，CEPH family）。来自每个个体的永生细胞系确保了 DNA 的永久供应，并且混合样品或非亲子关系样品早已通过分析许多标记而被排除掉。例如，1998 年 CHLC（人类连锁合作中心）图就是建立在使用 8325 个微卫星标记分析 8 个 CEPH 家系的结果的基础上，产生了 100 多万个基因型（Broman *et al.*,1998）。

13.4.3　多点疾病-标记定位

疾病-标记定位的出发点是标记的框架图。这是假定的，而且目标是将疾病基因定位至框架图的一个区间内。像 Linkmap（Linkage 软件包的一部分）或者 Genehunter（节 13.6.2）这样的程序能够在标记框架图上标出疾病基因座，计算在每个位置上系谱数据的总似然性。结果（图 13.8）是以图位置为对照的对数优势比曲线。这种方法也有益于排除定位：如果在区域内曲线停留在对数优势比为－2 以下，那么疾病基因座就排除在那个区域之外。

图 13.8 中明显的数量性状是极其谬误的。峰高主要取决于标记间精确的遗传距离，而对这些标记通常只是非常粗略地了解。而且，连锁程序中的作图函数（节 13.1.3）

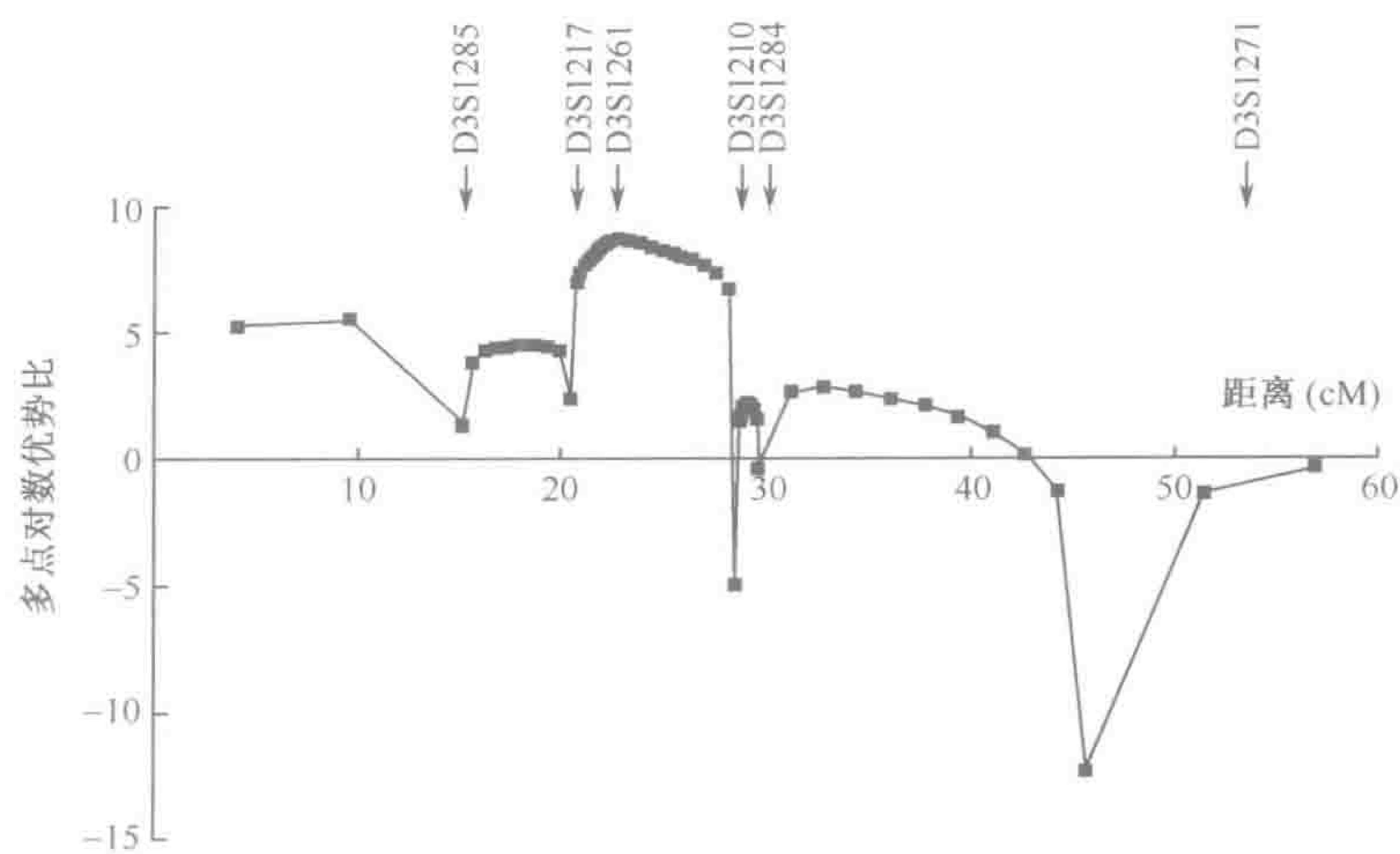


图 13.8　人类多点定位

水平轴为标记框架图而垂直轴为对数优势比，出自对一个 Waardenburg 综合征家系的分析。计算每一个可能的疾病基因座位置的对数优势比。在显示疾病重组体的位置附近，对数优势比强烈地降为负值。最高峰标志最可能的位置；支持这一位置的优势以最高峰超出其他竞争者/对手（位置）的程度来衡量。经 Nature Publishing Group 允许再引自 Hughes 等（1994），Nat. Genet. 7，509～512。

没有一个更接近于交叉分布（图 13.4）的真实复杂性。然而，如果标记图不是完全错误，最高峰代表最可能的位置这一点仍是真实的。如上所述，如果标记框架在物理上被锚定，那么接下来就是着手寻找候选区间内的 DNA 并鉴定疾病基因。

13.5　利用扩展的系谱和祖先单体型进行精细定位

定位的分辨率取决于减数分裂的数目——分析的减数分裂越多，缩小连锁区域的重组事件发生几率越大。大多数人类家庭的小规模严重地限制了家系研究中可达到的分辨率。然而，有时候扩展的家系结构可用于高分辨的定位。在一些群居人中，人们非常清楚他们的党派资格，并把他们自己看作是非常大的家系的一部分。即使在一些群居中人们限定他们家系感情为近亲，每个人在根本上还是相关的，并且有些时候在那些“无关”的人们中能够鉴定出共同祖先的染色体片段。常染色体隐性疾病适用于这样的分析，因为一个突变的等位基因能够传递许多代；对于大多数显性或 X 连锁疾病来说，其突变等位基因的转归太快，而在扩展的家系中不能共有（节 4.5.2）。假设一个人无法鉴定已定位的隐性疾病的携带者，定位将受限于从共同祖先遗传该疾病而亲缘关系疏远的受累者的数目。

13.5.1　同合性定位能够有效地在扩展的近亲家系中定位隐性疾病

同合性（autozygosity）是一专用名词，用来表示遗传自一个最近的共同祖先，具有传递一致性的标记的纯合性。近亲家系中患有罕见的隐性疾病的人在与疾病基因座连锁的标记处可能是同合性的。假设双亲是二级表亲：由于他们有共同的祖先，所以他们

必将共享 1/32 的基因，而孩子将只有 1/64 的基因座是同合性的。如果那个孩子一个特定的标记等位基因是纯合性的，那么这可能是由同合性引起的，或者是因为相同的等位基因的另一个拷贝已经独立地进入了该家系。等位基因在群体中越罕见，纯合性代表同合性的似然性越大。对于一个相当罕见的等位基因而言，二级表亲所生的单一的纯合性受累的孩子产生的对数优势比为 $\log_{10}(64)=1.8$。如果有两个其他的受累同胞，均在同一罕见的等位基因是纯合性的，那么对数优势比为 3.0 [$\log_{10}(64\times4\times4)$，即使同胞对于疾病而言是无亲缘关系的，同胞遗传相同的双亲单体型的概率为 1/4]。

因此，相当小的近亲家系能够产生有意义的对数优势比。如果能够发现具有多个相同隐性疾病受累个体由于近亲婚配而形成两个或更多的同胞群的家系，同合性定位则是一个特别有力的工具。在近亲婚配常见的中东国家可以找到合适的家系。这种方法已成功地应用于定位常染色体隐性耳聋基因（Guilford *et al.*,1994；图 13.9）。广泛的基因座异质性使隐性耳聋不可能在小核心家系标本中进行分析。

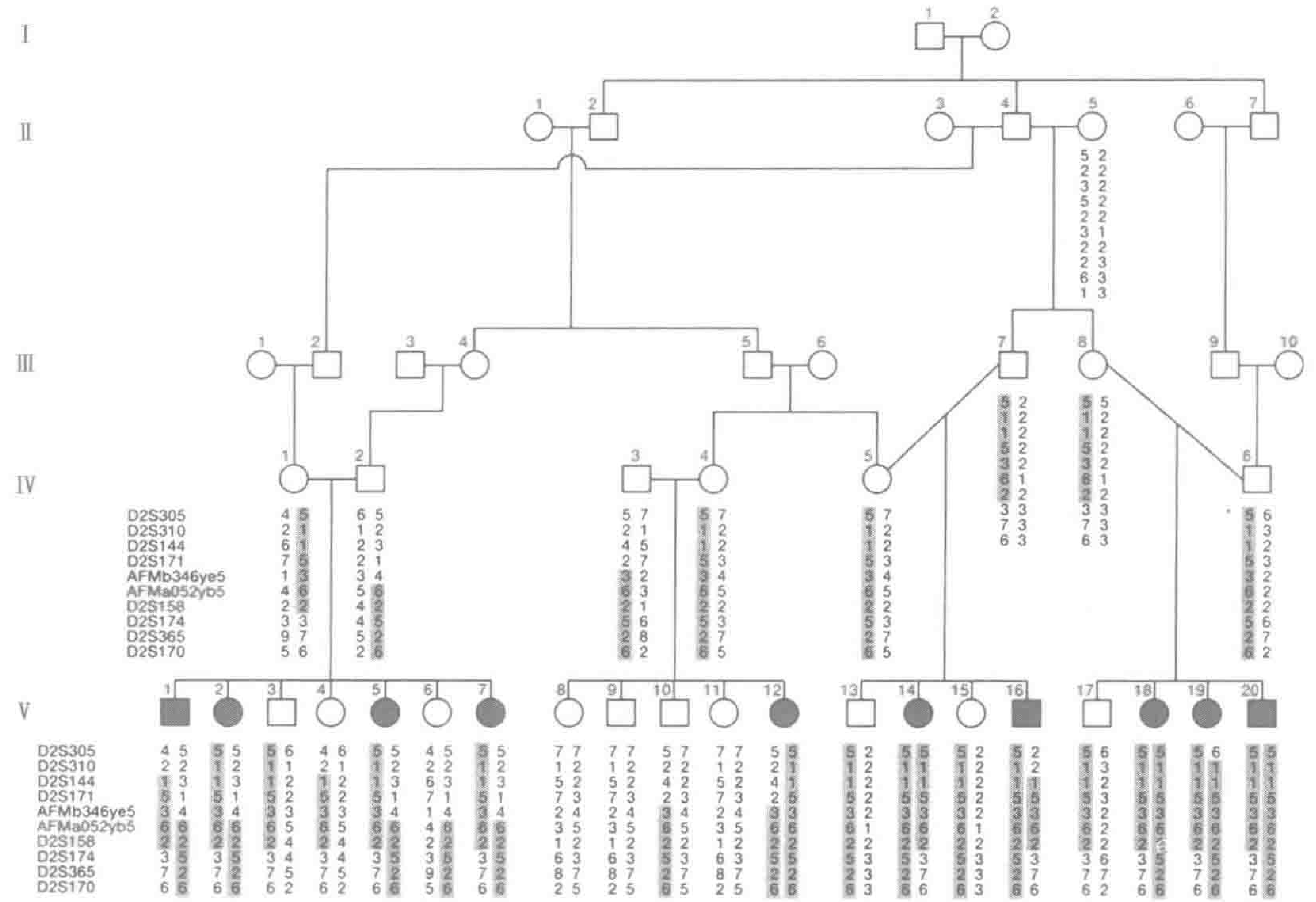


图 13.9 同合性定位

一个大的近亲繁殖的家系，有若干成员患有严重的常染色体隐性先天性耳聋（实心符号）的。黑色标示 2 号染色体上与耳聋一起分离的标记的单体型。所有受累的人在标记 AFMa052yb5 和 D2S158 是纯合性的，而未受累的人不是纯合性的。耳聋基因一定存在于这些标记两侧的两个标记（AFMb346ye5 和 *D2S174*）之间的某一处。经 Nature Publishing Group 允许再引自 Chaib 等（1996）. Hum. Molec. Genet. 5，155～158。

同合性在北欧人群中一个引人注目的应用使 Houwen 等（1994）仅利用了来自一个隔离的荷兰人村庄的四个受累个体（两个同胞和两个假定无关个体）就定位了罕见的隐性疾病——良性复发性肝内胆汁淤积。芬兰报道了类似的同合性应用。共享的祖先越

遥远，由于那个共同祖先而共享的基因组的比例越小，因而证实患者共享传递一致性的片段的意义越大。但同时，共同祖先越遥远，第二个独立的等位基因从外部进入这个家系的概率就越大，因此，无论对于疾病还是标记而言，纯合性代表同合性的可能性越小。就像 Houwen 等的研究一样，由于遥远的共同祖先，每件事都依赖于发现具有罕见隐性疾病并且其罕见的标记等位基因或（更可能的）单体型均为纯合性的人。Houwen 研究的效力看起来好像是不可思议的，但重要的是应记住这种方法学仅应用于这样的疾病和群体，其大多数受累的人是传自一个是携带者的共同祖先。

13.5.2 鉴定共享的祖先片段可对囊性纤维化和 Nijmegen 断裂综合征的基因座进行高分辨定位

囊性纤维化（CF）在家系结构比较适于同合性定位的非欧洲国家非常罕见，因此定位 CF 基因依赖于罕见的、不幸的、具有一名以上受累孩子的核心家系。利用这些家系，CF 基因定位于 7q31.2，但是，在利用了所有的重组体之后，候选区域仍旧非常大(这是在 20 世纪 80 年代末期，那时定位克隆是英雄们的工作)。当争论 CF 基因突变可能非常过时后（对杂合子不仅没有选择，而且可能是对他们有利的正性选择，节 4.5.2)，研究者着手从“无亲缘关系”的患者中鉴定 CF 染色体上共享的祖先染色体片段。重复地发现相同的标记等位基因单体型预示着共享。这种现象称为**连锁不平衡**(Linkage disequilibrium，LD)；更全面的描述见节 15.4。表 13.2 显示来自 CF 候选区域内两个标记的典型数据。非 CF 染色体表现出单体型的随机选择，但是 CF 染色体倾向于携带 X_1，K_2。其意义在于 LD 是一个非常短距离的现象（由于再发重组，共享祖先片段短)，因此，这暗示研究者难以找到的 CF 基因的确切位置。

表 13.2　囊性纤维化等位基因关联分析

标记等位基因	CF 染色体	正常染色体	标记等位基因	CF 染色体	正常染色体
X_1,K_1	3	49	X_2,K_1	8	70
X_1,K_2	147	19	X_2,K_2	8	25

数据来自在 114 个有囊性纤维化(CF)儿童的英国家系中 RFLP 标记 XV2.c(等位基因 X_1 和 X_2)和 KM19(等位基因 K_1 和 K_2)的分析。携带 CF 致病突变的染色体倾向于携带 XV2.c 的等位基因 X_1 和 KM19 的等位基因 K_2。数据来自 Ivinson *et al*.(1989)

一个比较近期克隆的控制 Nijmegen 断裂综合征（NBS；MIM 251260）的基因显示了同一原理更详细的应用。NBS 是一种非常罕见的常染色体隐性疾病，其特征为染色体断裂、生长迟滞、头小畸形、免疫缺陷和肿瘤易患性。可能的原因是 DNA 修复缺陷。在小核心家系内常规的连锁分析将 *NBS* 基因座定位至 8p21，但是在使用了所有的重组体之后，靶区域仍旧跨越了标记 *D8S271* 与 *D8S270* 之间 8Mb 的距离。接着对 51 名明显的无亲缘关系的患者及其双亲分析了跨越候选区域的一系列微卫星标记，产生了 102 种 NBS 单体型。其中，74 个单体型看起来像是一个斯拉夫起源的共同的祖先单体型的衍生物（图 13.10)。图 13.10 中最保守的区域包括标记 11 和标记 12，因此它标志 *NBS* 基因可能的位置。随后，从此位置克隆了一个编码新的蛋白质的基因，此基因在

NBS 患者中携带突变。正如所料，具有共同单体型的患者均具有相同的突变，而那些具有独立单体型的患者则具有独立的突变（Varon *et al.*,1998）。

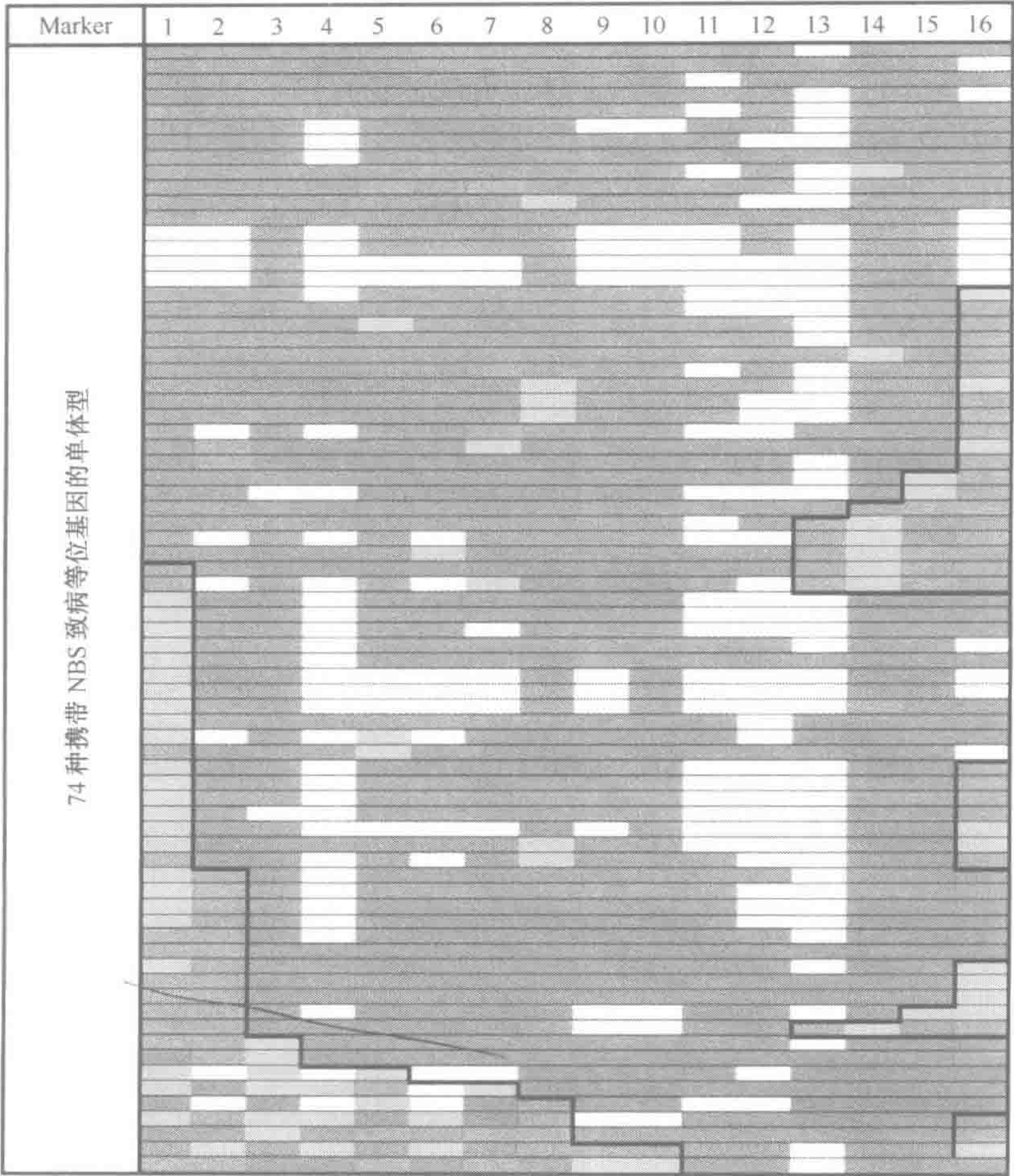


图 13.10　欧洲-Nijmegen 断裂综合征患者祖先的单体型

明显无亲缘关系的 NBS 患者通常看起来像已经从共同祖先遗传了 8p21 处染色体片段。利用 16 个标记确定了单体型，这 16 个标记按照染色体上的排列顺序分布于表的顶端。粉红色代表了与预测的祖先单体型一致的等位基因所在的位置。当非祖先等位基因与祖先等位基因仅 1bp 或 2bp 差异时，它们以黄色代表，可能由祖先等位基因突变衍生而来。以灰色代表的等位基因真正不同于祖先等位基因，并可能是重组的结果。空白代表了无数据的基因座。只有在基因座 11 和 2 是无重组体的等位基因（灰色），表明 *NBS* 基因定位于此处。数据出自 Varon 等（1998）。

LD 是尝试鉴定复杂疾病易感基因重要的工具，将在节 15.4 中详细讨论。

13.6　标准对数优势比分析不是毫无问题的

标准对数优势比分析对在 20Mb 片段内扫描基因组并定位一疾病基因是一种非常有力的方法，但是此方法陷入了困境。这些困难包括：

- 易出现误差
- 计算上能够分析何种系谱的限制
- 基因座异质性问题
- 最终可完成的分辨率的限制
- 需要设定精确的遗传模式，详细描述遗传方式、基因频率及每种基因型的外显率

13.6.1　基因分型和错误诊断的误差能够产生假重组体

由于标记的高度多态性，普通的误差诸如读胶错误、调换标本或者非血缘关系等通常将造成孩子既定的基因型与双亲的基因型不符。连锁分析程序将会停止直到这样的误差被纠正。引入可能的但却是错误的基因型的误差是更常见的问题，尤其是对某个人的疾病状态的错误诊断。这样的误差因引入假重组体而使遗传图的长度膨胀：如果一个孩子被赋予了错误的亲本等位基因，他将看起来就像是一个重组体。多基因座分析能够帮助解决此问题，因为假重组体表现为邻近双重重组体（图 13.11）。如我们在节 13.1.3 所见，干涉形成邻近双重重组体几乎不可能。当标记框架图制成后，常规误差校对通过忽略任何一个检验结果的方法检验该图能够缩短的程度（Broman *et al.*,1998）。显著地延长此图的结果（即增加重组体）是令人怀疑的。

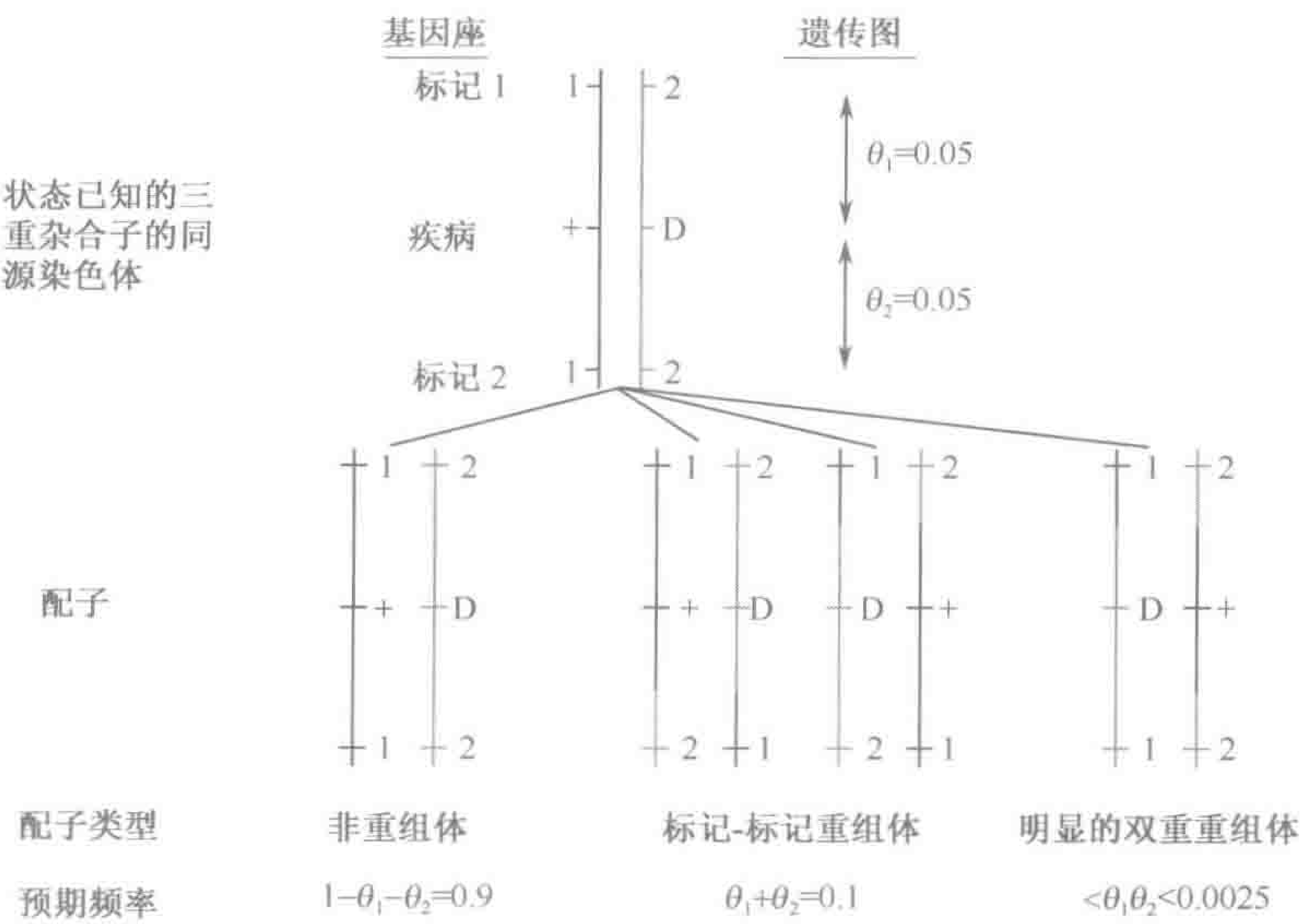


图 13.11　明显的双重重组体显示数据错误

由于干涉（节 13.1.3），相距 5cM 的标记形成真正的双重重组体的概率小，多小于 0.05×0.05＝0.0025。明显的双重重组体通常标志记录标记错误，临床误诊，或者基因座异质性即像此病例中致病基因不定位于基因座 D 而是基因组其他位置。某一基因突变或者生殖系嵌合体是比较罕见的原因。

标记框架图上标记顺序的误差常常令人头痛（单一重组体可能看起来像双重重组体），但在遗传图与物理序列相互校对后，这个问题已经消失了。

13.6.2 计算的困难限制了能够分析的系谱

如我们在节 13.3.2 所见，给定系谱数据和基因频率后，人类连锁分析依赖于计算机程序执行处理基因型概率分支树的运算。Liped 是第一个通常应用的程序，Mlink（Linkage 软件包的一部分）使用了相同的基本运算——Elston-Stewart 运算，但将其延伸至多点数据。Elston-Stewart 运算能够任意地处理大量系谱，但随着可能单体型（更多的等位基因和/或更多的基因座）数目的增加，计算的时间呈指数增加。这限制了 Mlink 分析多点数据的能力。另一个可供选择的运算，Lander-Green 运算，能够处理任何数目的基因座（随着基因座数目的增加，计算的时间呈线性增加），但对大量系谱存在记忆问题。这种运算在 Genehunter 程序（Kruglyak *et al.*，1996）和 Merlin 程序（Abecasis *et al.*，2002）中得以实现。这些程序尤其有益于适当大小的系谱进行全基因组扫描分析。

连锁分析的一般性理论在 Ott 的书（见进一步阅读）中作了极好的描述，而 Terwilliger 和 Ott 编写的书（进一步阅读）中记录了对于从事人类连锁分析的任何人都是必不可少的实际建议。

13.6.3 基因座异质性总是人类基因定位的一个陷阱

如我们在节 4.1.4 所见，几个没有连锁的基因的突变产生相同的临床表型是很常见的。如果在研究的家系标本中存在**基因座异质性**（locus heterogeneity），那么即使是拥有大量家系的一个显性疾病也难于定位。多年的合作研究表明结节硬化症是由 9q34 的 *TSC1* 基因座（MIM191100）和 16p13 的 *TSC2* 基因座（MIM 191092）两者中的任何一个突变引起的。隐性疾病因需要组合许多小家系而增加了困难。同合性定位（节 13.5.1）是这类病例的主要解决方法。

Genehunter、Homog 及相关程序（Terwilliger 和 Ott，进一步阅读）能够比较基因座同质性（所有家系定位至在检验中的位置）与基因座异质性（不连锁家庭的比例 α）两种假设的数据似然性，并给出 α 最大似然性的估计值。

13.6.4 减数分裂定位具有有限的分辨率

定位的分辨率取决于分析的减数分裂的数目。为此，人类家系是十分有限的——例如，CEPH 家系标本库（节 13.4.2）能够提供平均仅约 3Mb 的分辨率。一种解决办法是应用精子。男性可能有太少的孩子用于高分辨定位，但是他们有效地产生无限数目的精子。对于成对的、PCR 能够扩增的标记，各个精子能被划分为重组体或非重组体。这种方法不能用于定位疾病基因，但它使具有必要技能的研究者可以进行标记-标记定位，达到任何想得到的分辨率。Lien 等（2000）描述了一个典型的应用。通过计算低至 0.0001 的重组值（Jeffreys *et al.*，2001），证明了非常固定的**重组热点**（recombination hotspot）（图 13.5）的存在。当然，这仅能提供男性重组信息。

13.6.5 不符合孟德尔遗传的性状不适于用本章描述的方法进行定位

在本章描述的对数优势比分析方法需要一个精确的遗传模式。必须全部指定遗传方

式、基因频率和每种基因型的外显率。对于孟德尔性状，提供似乎可能的特征通常不是一个问题。外显率可能需要一些思考。如果没有考虑到非受累者可能是非外显的基因携带者，或者受累者可能是表型复制，那么这样的人将被划分为重组体。另一方面，如果外显率设置得太低，由于检验了一个不太精确的假说，检测连锁的效力将下降。然而，对于像糖尿病或精神分裂症这样的常见复杂病，问题更棘手。任何一种遗传模式都只不过是一种假说——我们对基因频率，任何易感等位基因的外显率，甚至遗传方式都没有真实的概念。这使得应用我们在本章中描述的方法研究这样的疾病变得非常不明智。尽管如此，鉴定复杂疾病易感性的遗传成分现在依然是人类遗传学研究的主要部分。人们尝试进行此项工作的方法是第 15 章的主题。

（邱广蓉　译）

进一步阅读

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd Edn. Johns Hopkins University Press, Baltimore, MD.

Terwilliger J, Ott J (1994) *Handbook for Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD.

参考文献

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**, 97–101.

Broman KW, Murray JC, Sheffield VC, White RL, Weber JL (1998) Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**, 861–869.

Broman KW, Weber JL (2000) Characterization of human crossover interference. *Am. J. Hum. Genet.* **66**, 1911–1926.

Chaib H, Place C, Salem N *et al.* (1996) A gene responsible for a sensorineural nonsyndromic recessive deafness maps to chromosome 2p22-23. *Hum. Mol. Genet.* **5**, 155–158.

Guilford P, Ben Arab S, Blanchard S, Levilliers J, Weissenbach J, Belkahia A, Petit C (1994) A non-syndrome form of neurosensory, recessive deafness maps to the pericentromeric region of chromosome 13q. *Nature Genet.* **6**, 24–28.

Houwen RHJ, Baharloo S, Blankenship K, Raeymaekers P, Juyn J, Sandkuijl LA, Freimer NB (1994). Genome screening by searching for shared segments: mapping a gene for benign recurrent intrahepatic cholestasis. *Nature Genet.* **8**, 380–386.

Hughes A, Newton VE, Liu XZ, Read AP (1994) A gene for Waardenburg syndrome Type 2 maps close to the human homologue of the microphthalmia gene at chromosome 3p12-p14.1. *Nature Genet.* **7**, 509–512.

Hultén MA, Lindsten J (1973) Cytogenetic aspects of human male meiosis. *Adv. Hum. Genet.* **4**, 327–387.

Hultén MA, Tease C (2003) Genetic maps: direct meiotic analysis. In: Cooper DN (ed) *Encyclopedia of the Human Genome*. Nature Publishing Group, London.

Ivinson AJ, Read AP, Harris R, Super M, Schwarz M, Clayton Smith J, Elles R (1989) Testing for cystic fibrosis using allelic association. *J. Med. Genet.* **26**, 426–430.

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222.

Kong A, Gudbjartsson DF, Sainz J *et al.* (2002) A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347–1363.

Lander ES, Schork NJ (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Lien S, Szyda J, Schechinger B, Rappold G, Arnheim N (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66**, 557–566.

Morton NE (1955) Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318.

Tease C, Hartshorne GM, Hultén, MA (2002) Patterns of meiotic recombination in human fetal oocytes. *Am. J. Hum. Genet.* **70**, 1469–1479.

Varon R, Vissinga C, Platzer M *et al.* (1998) Nibrin, a novel DNA double-stranded break repair protein, is mutated in Nijmegen Breakage Syndrome. *Cell* **93**, 467–476.

Wang DG, Fan JB, Siao CJ *et al.* (1998) Large-scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

第14章　鉴定人类致病基因

本章内容

本章更加准确但似欠生动的题目是“鉴定人类表型的遗传决定因素”。本文所描述的方法对于鉴定疾病或正常变异诸如红头发或红绿色盲的决定因素是同样适用的。不是所有可被鉴定的决定因素都必定是基因，即编码蛋白质的序列。已明确的是它们一定会影响表型，可通过某个间接途径影响蛋白质编码基因的表达水平或其mRNA的加工和稳定性。了解为什么一定的DNA序列变异体会引起特殊的表型是分子病理学的任务(16章)；在此，我们将讨论如何鉴定正确的变异体。

重要的是，不要被诸如“囊性纤维化基因”、“糖尿病基因”等这类词组所误导。许多人类基因是在研究它们突变所引起疾病时被第一次发现的，由于要经过很多年才能了解其正常功能，因此一直沿用这样的命名方式。然而，你不会将你的家用冰箱描述为“破坏冷冻食物的机器”。基因在细胞内发挥作用；假如作用没有完成或发生错误，结果可能导致疾病。

很少学科像鉴定人类致病基因那么快地提出。1980年以前，几乎没有人类基因被鉴定为致病基因。少数的成功例子包括已知生化基础并可纯化基因产物的少数疾病。20世纪80年代重组DNA技术的发展提供了新方法，有时却被给予毫无意义的称号“逆向遗传学”。鉴定致病基因的数量开始增加，但这些早期成功是艰难获得的英雄般的成果。随着用于连锁研究和突变筛查的PCR技术的发明，一切变得更加简单了。既然人类和其他基因组计划有效地获得了大量的资源，鉴定孟德尔疾病基因的能力几乎完全

取决于拥有合适的家系。鉴定常见复杂疾病相关因素易感性仍是极其困难的。

14.1　鉴定致病基因的原理和策略

有许多不同方法可以完成最终的鉴定（图 14.1），但所有的途径都集中于一个候选基因。通过某种方法鉴定一种候选基因；研究人员随后通过在该病的患者中筛查突变来验证这是致病基因的假设。

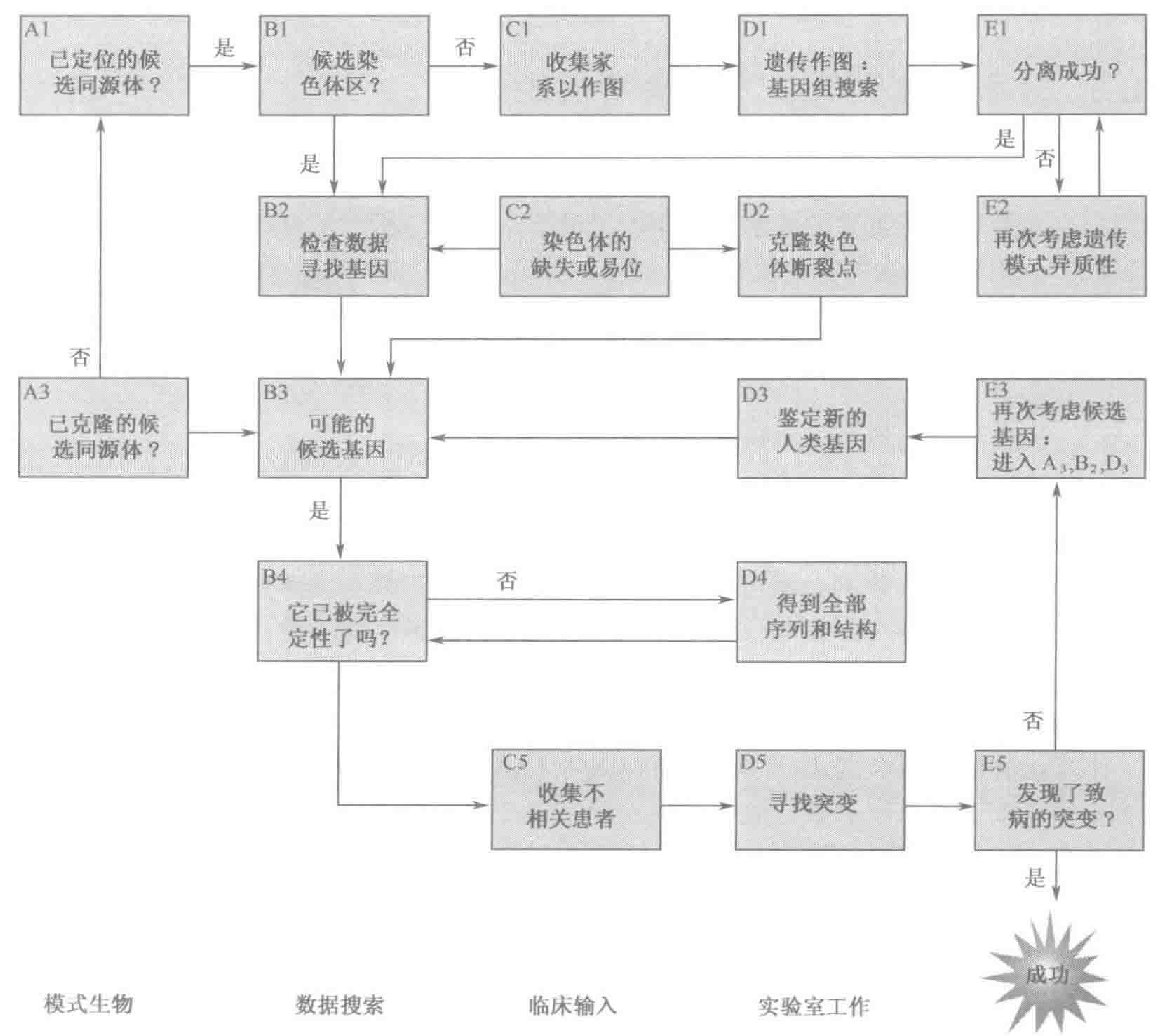


图 14.1　如何鉴定人类的致病基因

获得成功不只有一种方法，关键步骤是得到一个可能的候选基因，然后可以在患病人群中检测基因的突变。注意临床工作、实验室研究和计算机分析之间的相互作用。随着基因组计划积累的信息，搜寻数据变得越来越来重要。

候选基因可在缺乏染色体定位的情况下被鉴定（节 14.2），但更为常见的是，首先准确地找到一个候选染色体区，再从此区中鉴定候选基因（节 14.3）。既然我们有一个好的（尽管是不完整的）人类所有基因的目录，鉴定候选基因的任务变得意想不到的容

易，但从长长的候选基因表中不难看到筛选突变的工作仍然很艰巨。

定位信息使可能的候选基因的名单从人类全部大约 30 000 个基因减少到候选区内可能的 10～30 个。这是很重要的，因为无论多么努力地试图猜测可能的候选基因，目前我们这样做的能力都是很有限的。当一个致病基因一次又一次地被最终鉴定时，突变为什么会引起特定疾病仍将完全是个谜。为什么涉及 RNA 从细胞核转运至细胞质的 FMR1 蛋白质的功能丧失引起智力低下和大睾丸（脆性 X 综合征 MIM 309550），而 TATA 结合蛋白的特定突变（节 1.3.4）引起 SCA17 脊髓小脑共济失调（表 16.6）？

14.2 不依赖定位的鉴定致病基因的策略

历史上第一个致病基因是通过不依赖定位的方法鉴定的，只不过因为不存在相关的作图信息及可用于产生它的技术。在那种情况下，必须根据对基因产物：镰状细胞病的 β 珠蛋白，苯丙酮尿症的苯丙氨酸羟化酶等的认识提出候选基因。直至今天，生化或细胞生物学方面的研究可以鉴定未知基因的蛋白质产物。某些方法需要从蛋白质转移到 DNA。

14.2.1 通过已知蛋白质产物鉴定致病基因

现代的蛋白质组学技术通过质谱分析（Mann *et al*.,2001 和节 19.4.2）和化学微序列测定（Bartlett，2001），可以鉴定或部分测序即使微量的蛋白质。如果可以得到编码那些氨基酸的 cDNA 序列，就可以合成寡核苷酸探针用于筛选文库以找回 cDNA。问题在于密码子的简并性——大部分氨基酸可以由几个密码子中的任何一个编码。探针应是一种**简并寡核苷酸**（degenerate oligonucleotide)，是所有可能序列的混合物，并与氨基酸的某一部分序列相匹配，且此段序列的可能前突变次数不会很大。由于在混合物中只有一个寡核苷酸会符合真正的序列，所以减少不同探针的数量，对增加鉴定正确靶序列的概率很重要。色氨酸和蛋氨酸因为各自只有唯一密码子，因此在此非常有帮助。而应尽量避免有 6 个密码子的精氨酸、亮氨酸和丝氨酸。

用一个简并寡核苷酸探针筛选文库是令人厌烦的，因为杂交条件会对结果产生很大影响。更加快捷一点的选择是使用部分简并的寡核苷酸作为 PCR 引物。将靶 cDNA 连到载体，并使用一个载体特异性引物和一个简并蛋白特异性引物，可以减少可能的前突变次数。然而，要获得所希望的 PCR 产物，而不是没有产物或大量不相关的产物，需要好运气。

另一种途径是，如果存在微量的可用蛋白质，则可产生此蛋白质的抗体，并借此方法发现基因。回顾 1982 年，编码苯丙氨酸羟化酶的 mRNA 是通过免疫沉淀在多聚核糖体中被发现的，多聚核糖体在无细胞体系里合成蛋白质（Robson *et al*.,1982）。目前通过克隆集中的 cDNA 到一表达载体可以构建 **cDNA 表达文库**（cDNA expression library)（节 5.6.1）。含有携带所需基因克隆的宿主细胞，可以产生蛋白质，或至少部分蛋白质，并且可以用适当的抗体在文库中筛选菌落滤膜后进行鉴定。这里，每项工作都取决于抗体的特异性，而且希望此蛋白质不会对宿主细胞产生毒性。**噬菌体展示**

(phage display)（节 5.6.2）提供了另一种方法。

14.2.2　通过动物模型鉴定致病基因

许多人类的致病基因是在动物模型的帮助下——但几乎总是在找到定位信息之后被鉴定的。也许小鼠的突变体和表型相似的人类疾病定位在染色体上的相应位点［使用**牛津网格**（oxford grid），见图 14.7］。然后，假如克隆了小鼠的基因，则其人类同源体就成为自然的候选基因。或者，在小鼠中可鉴定一种致病基因，然后其人类同源体被分离；可通过荧光原位杂交将其定位（节 2.4.2），它就成为定位于此位点任一相关疾病的候选基因。2 型 Waardenburg 综合征（MIM 193510；Hughes *et al.*,1994）的致病基因 *MITF* 就是这样鉴定的。

在没有任何定位信息表明这些检测适合患者之前，把在动物模型中鉴定的基因直接在人类患者中进行检测是罕见的，但 *SOX10* 是一个这样进行检测的例子。该基因是通过对小鼠显性巨结肠（Dom）突变的艰辛定位克隆得以鉴定的。Dom 小鼠是一个人类先天性巨结肠症的长期研究模型（节 15.6.2）。患者合并患有先天性巨结肠症，色素异常及听力丧失（Waardenburg 综合征Ⅳ型或 WS4，MIM 277580），与小鼠特别相似。WS4 非常少见，而通常发生在太小而难以定位的家系中，所以在先前没有任何关于它们疾病可以定位在哪里的信息的情况下，在一组 WS4 患者中检测了 *SOX10* 突变。尽管不是在所有患者中都发现突变，但这个赌注在 *SOX10* 突变被发现时偿还清了（Pingault *et al.*,1998）。

14.2.3　应用不依赖定位的 DNA 序列信息鉴定致病基因

研究者在考虑什么疾病可由某特定已知基因的突变引起时常用此种方法。不依赖定位的候选基因也可通过表达阵列实验发现，在此实验中，对比患者与对照组的 mRNA 在该疾病中表达发生了改变进而产生一个基因表。

不依赖定位的 DNA 序列信息的一个有趣的应用，是用于克隆具有新的三核苷酸扩展重复基因。正如节 16.6.4 所述，扩展的三核苷酸重复引起几种遗传性神经疾病。通常，这些疾病表现为早现遗传——在较早的年龄发病，并在连续的世代中严重性不断增强。假如研究中的某疾病具有这些特征中的任何一种，应有必要从受累患者的 DNA 中筛查三核苷酸重复。Schalling 等（1993）的重复扩展检测方法可以检测到受累患者未被分段的基因组 DNA 中的重复扩展，而且该方法已经发展到用于克隆任何检测到的重复扩展（Koob *et al.*,1998）。该方法被应用于完全不依赖定位方式鉴定一类脊髓小脑共济失调综合征（SCA8）中的新的重复扩展。

14.3　定位克隆

在定位克隆时，仅知道致病基因大概的染色体位置，就可以鉴定致病基因。首个成功的应用是鉴定 X 连锁慢性肉芽肿病的基因（Royer-Pokora *et al.*,1985）。定位克隆方法的一个主要实验床（test-bed）是 Duchenne 肌营养不良（DMD，MIM 310200）。对

受累肌肉的病理改变逐年的仔细研究无法揭示 DMD 的生化基础。在 20 世纪 80 年代早期，几个研究小组用不同方法竞相克隆了 DMD 基因。这些研究小组开拓性的工作，战胜了极大的技术困难，克隆了这个史无前例的基因。这些工作也为后续的大部分定位克隆工作予以重要的启示。Worton 和 Thompson（1988）对此项工作做了很好的综述。

1986 年，此项工作成功地完成标志着壮丽的人类分子遗传学新纪元的开始。人类潜在的重要疾病基因一个接一个地被陆续分离，如囊性纤维化、亨廷顿病、成人多囊肾病及家族性结肠癌。图 14.2 显示了定位克隆的逻辑过程。然而在当今的标记图、克隆和序列分析技术应用之前，定位克隆是一项难度极大的工作。到 1995 年，通过此方法仅鉴定了约 50 种遗传病的基因。一位研究者在图 14.3 中总结了定位克隆的艰难特性。

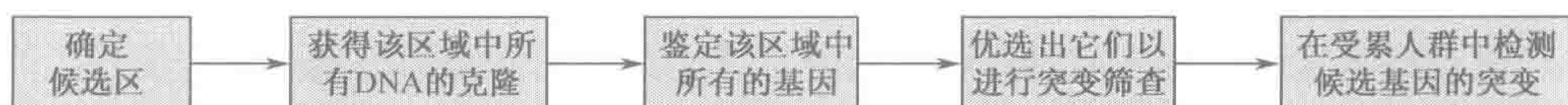


图 14.2 定位克隆的逻辑过程

图表显示了对位置克隆的推理过程；然而直至今天，序列数据库和高分辨标记图的应用，才使研究者能尝试所有捷径，以减少纯位置克隆的工作量。

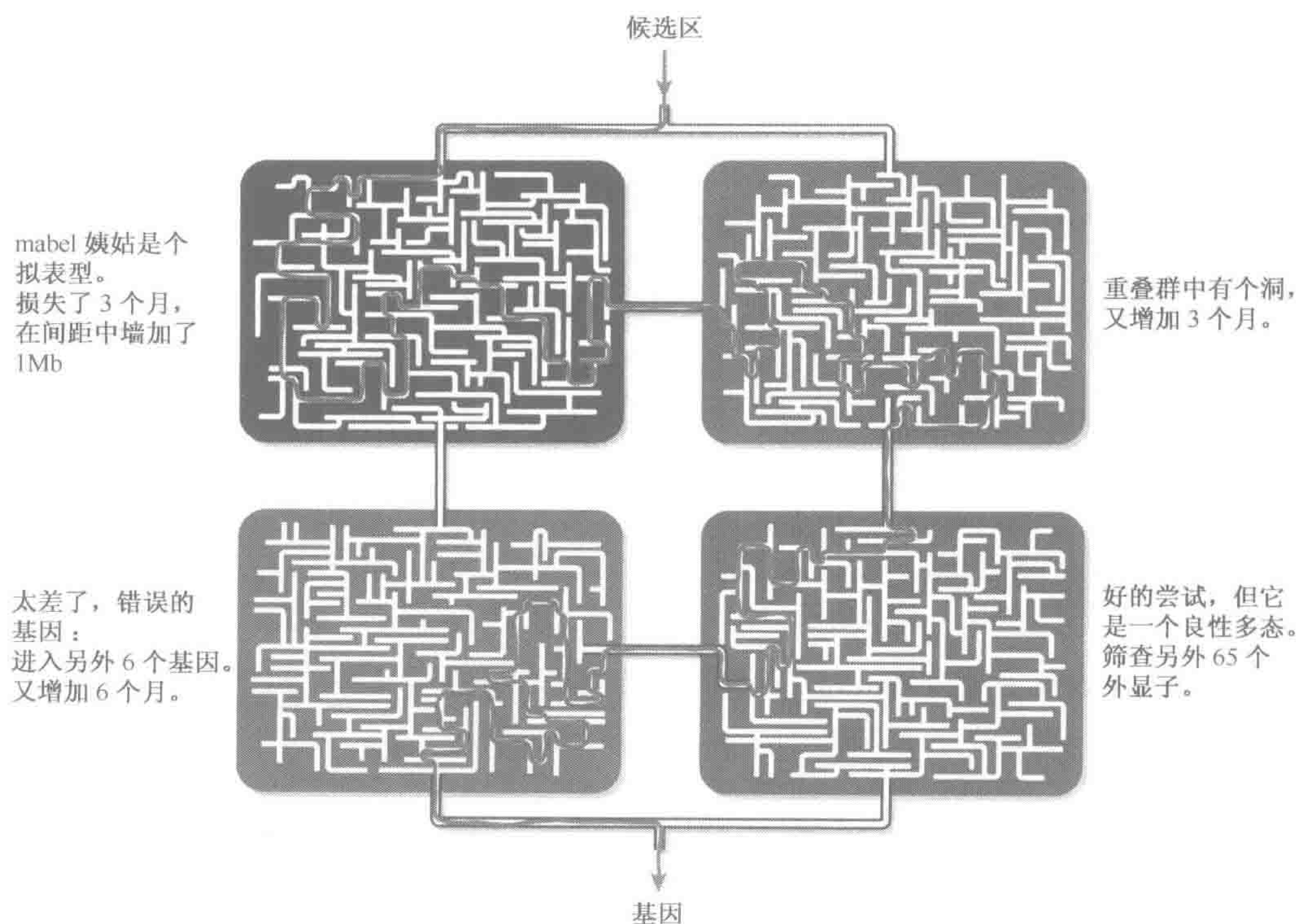


图 14.3 从候选区域到基因的艰难途径

一位研究者眼中艰难的位置克隆工作。University of Iowa，Dr. Richard Smith 惠允。

14.3.1　第一步是确定尽量小的候选区

定位克隆的艰难主要取决于候选区范围的大小，所以首要优先的是尽可能缩小范围。对于孟德尔疾病来说，应用控制减数分裂次数可以达到以上目的。当两个邻近标记之间最后的重组体被定位后，也就达到了**分辨的极限**（limit of resolution)。这是由单体型的调查得出的，而非计算机分析的结果（图 14.4)。应用经验规律，遵循 1cM＝1Mb（节 13.1.5)，一个家系收集了 100 个可提供信息的减数分裂，就可将一种孟德尔疾病定位于大约 1Mb 的候选区内。

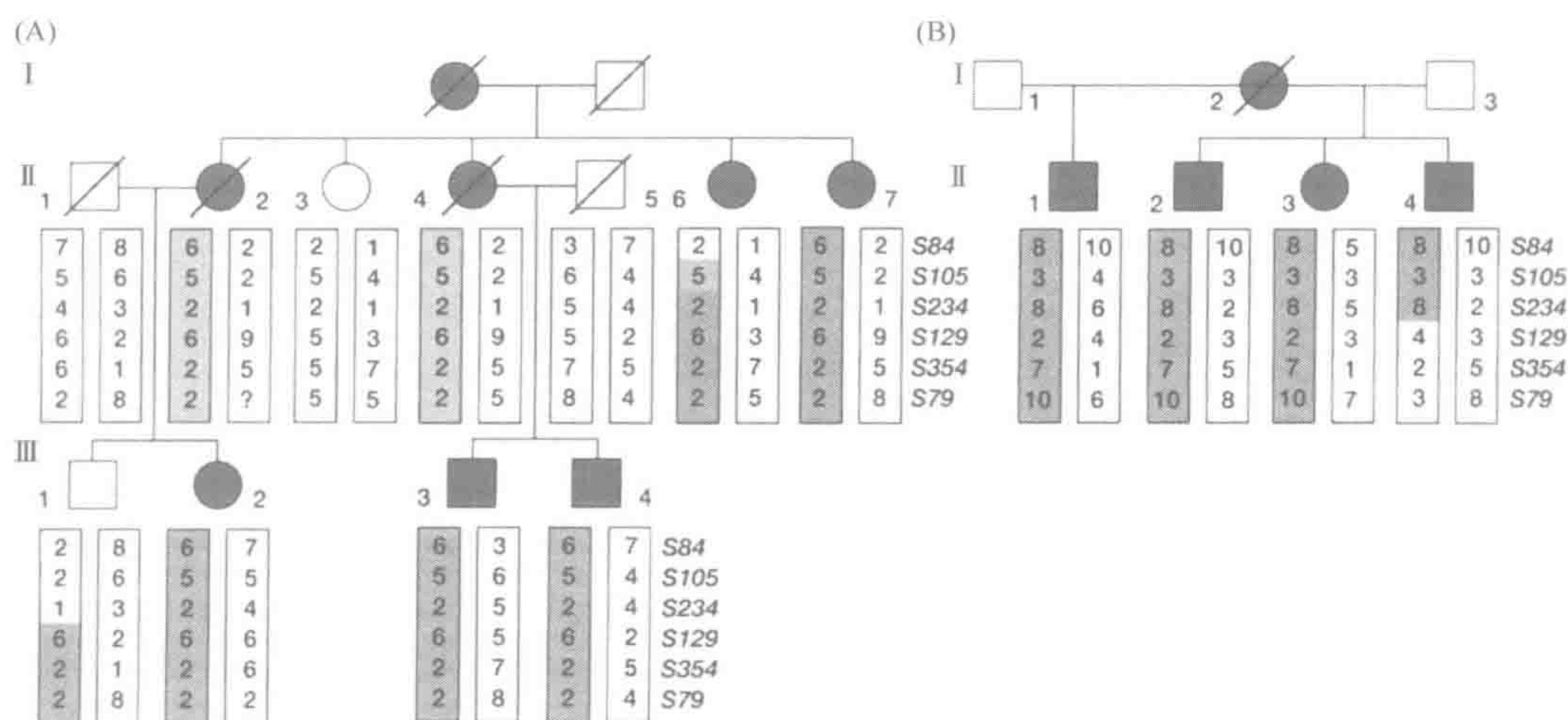


图 14.4　通过单体型检测来确定候选区

以上两个系谱显示，一种显性遗传性皮肤病 Darier-白化病（MIM 124 200）以前被定位于 12q。与疾病分离的 12q 标记单体型由亮色表示，灰色框标记去世的人群的推测的单体型。系谱 A 中，个体Ⅱ-6 重组体将致病基因定位于 *D12S84* 的远端；*D12S105* 无法提供信息，因为比较Ⅱ-3 和Ⅱ-7 的基因型，可看出Ⅰ-1 是明显的等位基因 5 的纯合子。重组体Ⅲ-1 提示致病基因定位于 *D12S129* 附近，但这也需要证实，因为这种解释依赖于对Ⅱ-1 和Ⅱ-2 基因型的正确推断和确定Ⅲ-1 不是非外显的基因携带者。系谱 B 中，Ⅱ-4 重组体提供了证明。此综合数据将 Darier 基因定位于 *D12S84* 和 *D12S129* 之间的区域。经授权，重绘自 Carter 等（1994）Genomics 24，378～382 © 1994 Elsevier。

当应用单个重组体确定候选区边界时，考虑可能的错误来源是很重要的（节 13.6.1)。谨慎的临床诊断是必要的。出现在明确受累人群中关键的重组体更值得可信——然而，非受累个体可能是非外显的基因携带者。**明显的双重组体**（apparent double recombinant）是极其可疑的，有时，尽管有好的阳性 lod 值，但用所有标记试验后，却似乎是重组体。这通常表明被研究的家系之一不定位在该区域内。另外，或许标记在遗传图上的排序不正确。

对于非孟德尔的表型，连锁分析远不够精确，典型的候选区是 20cM 或更大（15 章)。如果没有进一步的线索，这个区域太大而难以寻找，在此，强调应用**连锁不平衡**（linkage disequilibrium）来缩小寻找范围（节 15.4)。即使是孟德尔疾病，连锁不平衡也是精细定位非常有用的工具，正如我们看到的囊性纤维化和 Nijmegen 断裂综合征

（节 13.5.2）。

14.3.2　必须构建跨越候选区的克隆重叠群

重叠群构建技术如框 8.5 所述。在早期，建立重叠群是主要的费力工作。描述鉴定囊性纤维化致病基因的文章（Rommens *et al*.，1989）总结了这个早期研究中可能给人印象最为深刻的例子。应用文库杂交筛选鉴定连续重叠克隆［**"染色体步查"**（chromosome walking）］是痛苦漫长的，而后应用小片段插入的噬菌体和黏粒文库，和现在过时的染色体跳跃技术（Poustka *et al*.，1987）作补充。尽管在完全信赖这些数据之前，常有必要进行检查，仍然可以从人类基因组序列数据库中下载得到已经构建好的重叠群。

14.3.3　转录物图确定候选区内的所有基因

构建重叠群之后，下一步的工作是对其中的全部基因编目。随着我们对人类基因组知识的增长，我们研究的任何一种疾病由已知基因突变引发的可能性日益增加。**基因组浏览器**（genome browser），如 Ensembl（http://www.ensembl.org）或 Santa Cruz 浏览器（http://gemome.cse.ucsc.edu），用于显示和分析候选区内所有明确的及可能的基因（图 14.5；也见节 8.3.6 和 Wolfsberg *et al*.，进一步阅读）。虽然这些方法给人以深刻的印象，但重要的是不能完全依赖于它们。这些尖端高新技术需要用于辅助思考，而不能代替思考。它们必须得到用联合计算机和实验工作获取的对该区域有深度的第一手资料作为补充。Reymond 等（2002）的文章中说明这一类额外分析的方法是可行的。

图 14.5　使用基因组浏览器来列举候选区域内的基因

这是 Ensembl 基因组浏览器的一部分界面，显示出在染色体 6p21.2 中长 1Mb 区域内确定的和预测的基因。点击一基因会显示关于此基因序列、结构及同源性等信息。该系统的更多细节，详见节 8.3.6。

计算机研究者们试图从数据库中强求比基因组浏览器程序能提取更多的数据。基因查找程序不善于寻找小的外显子，这些外显子有罕见的剪切位点或是密码子偏斜，或是具有少见长的 5′或 3′端非翻译区域。表达序列标签（EST）由剪接的小外显子组成，可能无法与它们的基因组序列相匹配。进一步的计算机分析可集中在检测同源性上。比较

人和小鼠的序列会发现另外的保守序列，提示可能具有某种功能。可能的种间同源体或种内同源体在应用全自动基因组注释时弱同源性可丢失。一种方向更明确的，以假设推动的搜索会给基因功能提供重要的指示。

实验室研究集中在双检查序列组合中的错误，如亚克隆的错误排序或基因组合中的错误，诸如外显子的丢失、假外显子和基因的剪切或连接等。与基因组序列不同部分相匹配的引物用于检查所产生的预期大小的产物。扩增失败提示基因组序列的错误组合。应用来自预期的不同外显子的引物的 RT-PCR，可以检查预测产物是否能被完全扩增，假如可以，它是否包含预期的中间的外显子。毋庸置疑，可以发现另外的剪接体或额外的外显子。5′-RACE（节 7.2.3）可以用于试图延伸基因组序列，特别是如果在外显子上游最远端没有好的起始密码子（如在 Kozak 一致序列中的 ATG，节 1.5.1）。当然，如果对预测基因的表达模式一无所知的话，扩增失败可能仅表示检查的是错误的组织或 cDNA 文库。然而，如果预测的基因序列的任何一种出现在 EST，就会有该 EST 从哪一个 cDNA 文库中分离出来的信息。邻近的基因沿相同的方向转录，可以应用 RT-PCR 实验来检查，是否它们实际上可能都是同一基因的一部分。

除了对数据注释进行实验检测外，可以**直接搜索转录物**（direct search for transcript）。节 7.2 中详细阐述了在基因组重叠群中鉴定未知转录序列的一般方法，框 14.1 中也有简要总结。直到最近仍没有数据库可检查，所以这些方法形成了第一代转录物作图。

无论何时筛选 cDNA 文库，都会产生使用哪个文库的问题。通常研究中的疾病的病理学提示了特异的调查方向。因此，当研究一种神经肌肉疾病时，从筛选肌肉 cDNA 文库开始是有道理的。然而，有病理表现的组织不一定是表达最强的，因此，当一个文库筛选失败时，筛选其他文库总是值得的。胎儿脑组织是常用的选择，因为它含有数量特多的表达序列。

框 14.1　转录物作图：增补在基因组克隆内鉴定表达序列数据分析的实验室方法

转录物作图的方法在节 7.2 中已有阐述。总之，它们包括：

- cDNA 文库筛选，用于探查候选区域内基因组克隆；
- cDNA 选择，用于源自候选区域 cDNA 的超灵敏检测（图 7.11）；
- 外显子捕获，由功能性剪接信号寻找基因组旁侧的序列（图 7.10）；
- 动物印迹，寻找进化的保守序列（图 7.9）；
- 鉴定 CpG 岛，寻找常位于靠近基因的低甲基化 DNA 的区域（框 9.3）。

14.3.4　候选区内的基因必须先进行突变检测

从定位在候选区内的基因名录中，人们应该寻找一个显示**适当表达**（appropriate expression）和/或**适当功能**（appropriate function）的基因。除此之外或附加的，如下所述，人们应寻找已知具有适当表达或功能，或具有相应表型突变体的其他人类或非人类基因的**同源性**（homology）。

适当的表达模式

一个好的候选基因应该具有与疾病表型一致的表达模式。表达不必仅严格限于受累组织，因为有许多广泛表达的基因引起组织特异性疾病的例子（节 16.7.1），但候选基因应该至少在可见病理改变之时之处表达。例如神经管缺陷可能与人类胚胎发育的第三和第四周时表达的基因相关，发生在神经胚形成之前不久或过程中。候选基因的表达可以通过 RT-PCR、Northern 印迹或基因表达系列分析（SAGE，节 19.3.2）检测。大部分前期工作可在数据库（dbEST 或 SAGE 数据库，都可进入 NCBI 首页 http://www.ncbi.nlm.nih.gov/）中而不是在实验室中完成。在组织切片上对 mRNA 原位杂交（节 6.3.4）或使用标记抗体的免疫组化来提供最详细的表达模式图像。通常在小鼠组织，特别是其胚胎阶段的组织进行研究。通常，推测人与小鼠会具有相似的表达模式的证明不总是正确的，同时已建立人类胚胎阶段的切片资料中心，使之在必要阶段进行关于人类胚胎的同等分析。

适当的功能

当候选区内的基因功能已知时，它是否为此疾病好的候选基因是显而易见的——视紫质和纤维蛋白原（节 14.6.4）提供了例子。作为新基因，序列分析常会为研究其功能提供线索：可以鉴定跨膜区域，酪氨酸激酶基序等。考虑到疾病的病理，这些可能足以使一个基因优先成为候选基因。例如，已知离子转运对内耳功能十分关键，因此，离子通道基因自然就成为定位克隆耳聋基因的候选基因。

候选基因也可根据功能密切相关并涉及相似疾病的已知基因的提示来确立。基因之间可以通过编码一种受体及其配基或者其他在新陈代谢或发育通路中相互作用的组件建立而联系。例如，某些参与先天性巨结肠症的基因就是应用此种逻辑被鉴定的，正如节 15.6.2 中所叙述的。

相关的种内同源的（人类）基因同源性

有时，候选区内的基因被证明是一个已知基因（人类的种内同源体或其他物种的种间同源体）的密切同源体。假如该同源基因的突变引起一种相关的表型，该新基因则会成为备受关注的候选基因。例如，当原纤维蛋白被鉴定为在 Marfan 综合征中发生突变的基因时（节 14.6.4），一种种内同源基因 *FBN2* 被定位于 5q。一种相关的情况，先天性挛缩性细长趾（CCA，MIM 121050）定位在 5q 同一区域。不久，就在 CCA 患者中证明了 *FBN2* 突变（Putnam *et al.*，1995）。

相关种间同源的（模式生物）基因的同源性

在过去十年中，结构和功能的同源性在即使关系很远的相关物种间能延伸多远也已变得日益清晰。事实上小鼠每一个基因都有一个精确的人类副本，对于其他很少深入研究的哺乳动物物种，也同样几乎是正确的。更令人惊奇的是，人类基因与斑马鱼、果蝇、秀丽新小杆线虫甚至是酵母的基因都可以检测到广泛的同源性。因此在众多人类基因中，优先考虑候选基因的一个非常有效的方法是，查看这些已经研究得很清楚的模式

生物中关于同源基因是哪些，如节 19.2 中所述。这些数据可能包括表达模式和突变表型。Steinmetz 等（2002）阐述了为寻找人类潜在的致病基因，而对酵母突变株的系统筛查。对于这种研究，小鼠是特别有用的，它们的应用在下文中有更详细的描述。

除了更多的基因序列，通路也常是高度保守的，因此，应用对果蝇或酵母的发育或控制通路的知识，可以预测人类通路可能的工作方式——尽管哺乳动物常有几种平行的路径对应于低等生物的单一路径。相反，突变体表型则较少会是密切相对应的。一个显著的例子是果蝇的无翅突变。人类的一种 *Lhx2* 基因可以弥补此种突变带来的功能上的缺陷，于是果蝇长出了正常的翅膀（图 14.6）。实际上我们也有与果蝇一致的发育通路，但很明显，我们是为了不同的目的而用它。鳃-耳-肾综合征（节 14.6.3）是另一个例子。

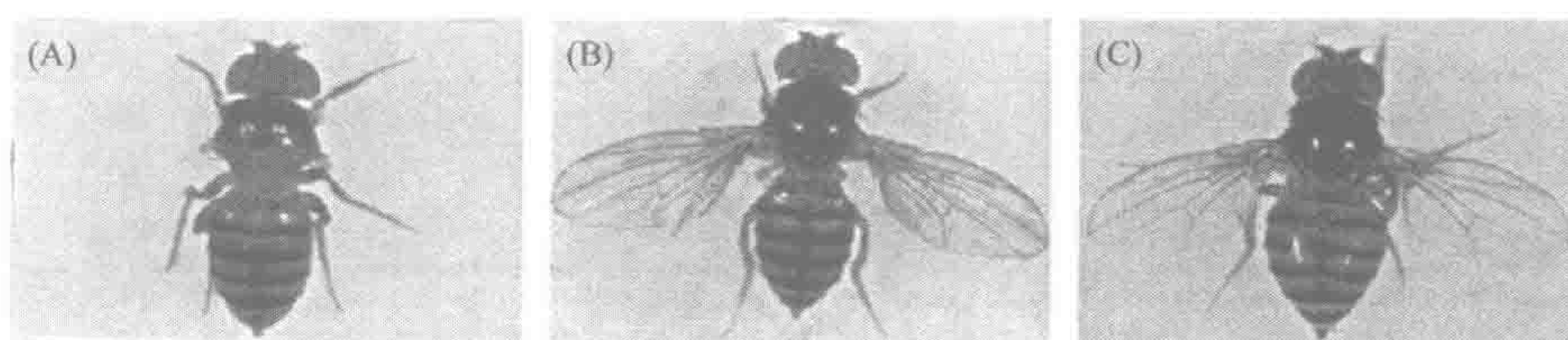


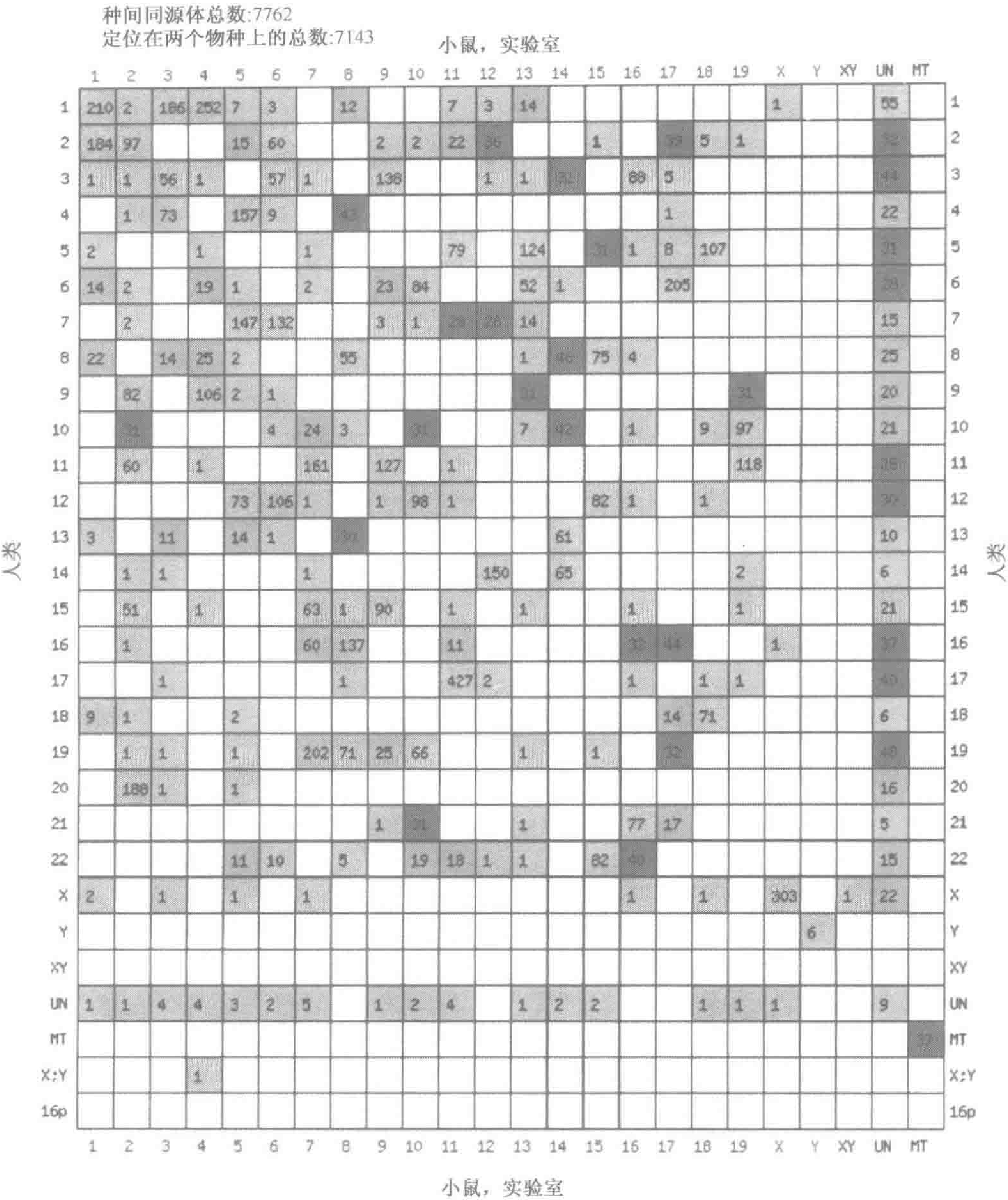
图 14.6 人类有使果蝇长出翅膀的基因

无翅突变的果蝇（A）可以被野生型果蝇基因（B）或人类 *Lhx2* 基因（C）所矫正。经授权，摘自 Rincón-Lima *et al*.（1999）Proc. Natl. Acd. Sci. *USA* 96，2165～2170。© 1999 National Academy of Sciences，USA。

14.3.5 小鼠突变体的特异相关性

人类-小鼠表型的同源性给鉴定人类致病基因提供了特别有价值的线索，有几点理由：

- 系统的诱变计划正产生很大数量的小鼠突变体（Justice，2000；Brown and Balling，2001）；
- 与人类、果蝇或线虫相比，种间同源基因突变在人类和小鼠更可能产生相似的表型。然而相似性可能不像所期望的那么密切（节 20.4.6）；
- 小鼠的表型信息常容易地翻译成定位的候选基因信息。回交作图（框 14.2）可以在小鼠中快速准确地作图。一旦已知一种感兴趣的基因在小鼠或人染色体上的定位，通常（尽管不总是）就可以预测此基因在其他物种中的位置。图 14.7（另见彩图）表示在小鼠和人类染色体位置间的大体相似性，这是基于两物种中都已定位的种间同源基因。人类与小鼠基因组序列的交叉匹配提供了人类与小鼠染色体关系的详细图谱（Gregory *et al*.,2002；图 14.8）；
- 外显子序列和外显子一内含子结构通常在人类与小鼠的种间同源基因间通常是高度保守的。这意味着一旦分离出人类或小鼠基因，可以设计探针或引物筛查来自其他物种的 DNA 文库，以鉴定种间同源基因；
- 一旦候选基因在人类中被鉴定，可以建立小鼠的突变体，用以进行功能分析。我们有能力在生物体内制造全部或有条件的基因敲除及制造特异性突变，这使得小鼠成为研究人类基因功能的非常有力的工具。



种间同源体总数:7762
定位在两个物种上的总数:7143

小鼠，实验室（列）；人类（行）

人类	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT
1	210	2	186	252	7	3		12			7	3	14							1			55	
2	184	97			15	60			2	2	22	36			1		39	5	1				32	
3	1	1	56	1		57	1		138			1	1	32		88	5						44	
4		1	73		157	9		43									1						22	
5	2			1			1				79		124		31	1	8	107					31	
6	14	2		19	1		2		23	84			52	1			205						28	
7		2			147	132			3	1	28	28	14										15	
8	22		14	25	2			55					1	46	75	4							25	
9		82		106	2	1							31						31				20	
10		31				4	24	3		31			7	42		1		9	97				21	
11		60		1			161		127		1								118				26	
12					73	106	1		1	98	1				82	1		1					30	
13	3		11		14	1		30						61									10	
14		1	1				1					150		65					2				6	
15		51		1			63	1	90		1		1			1			1				21	
16		1					60	137			11					32	44			1			37	
17			1					1			427	2				1		1	1				40	
18	9	1			2												14	71					6	
19		1	1		1		202	71	25	66			1		1		32						48	
20		188	1		1																		16	
21									1	31			1			77	17						5	
22					11	10		5		19	18	1	1		82	40							15	
X	2		1		1		1									1		1		303		1	22	
Y																					6			
XY																								
UN	1	1	4	4	3	2	5		1	2	4		1	2	2			1	1	1			9	
MT																								37
X;Y				1																				
16p																								

图 14.7　人类与小鼠的基因图间的保守同线性

牛津网格总结了人类和小鼠染色体间的相互关系。根据定位的种间同源体数量，格子被标记上颜色。很明显是非随机分布的。如果小鼠位置是已知的，通常可以预测人类基因的位置，反之亦然。在已知小鼠基因定位的情况下可以预测人类基因定位。此图只是一种概况，详细信息包含在数据库http://www.ncbi.nlm.nih.gov/Omim/Homology（DeBry and Seldin，1996）中。图表授权自 Mouse Genome Database，Mouse Genome Informatics，The Jackson Laboratory，Bar Harbor，Maine（http://www.informatics.jax.org）。

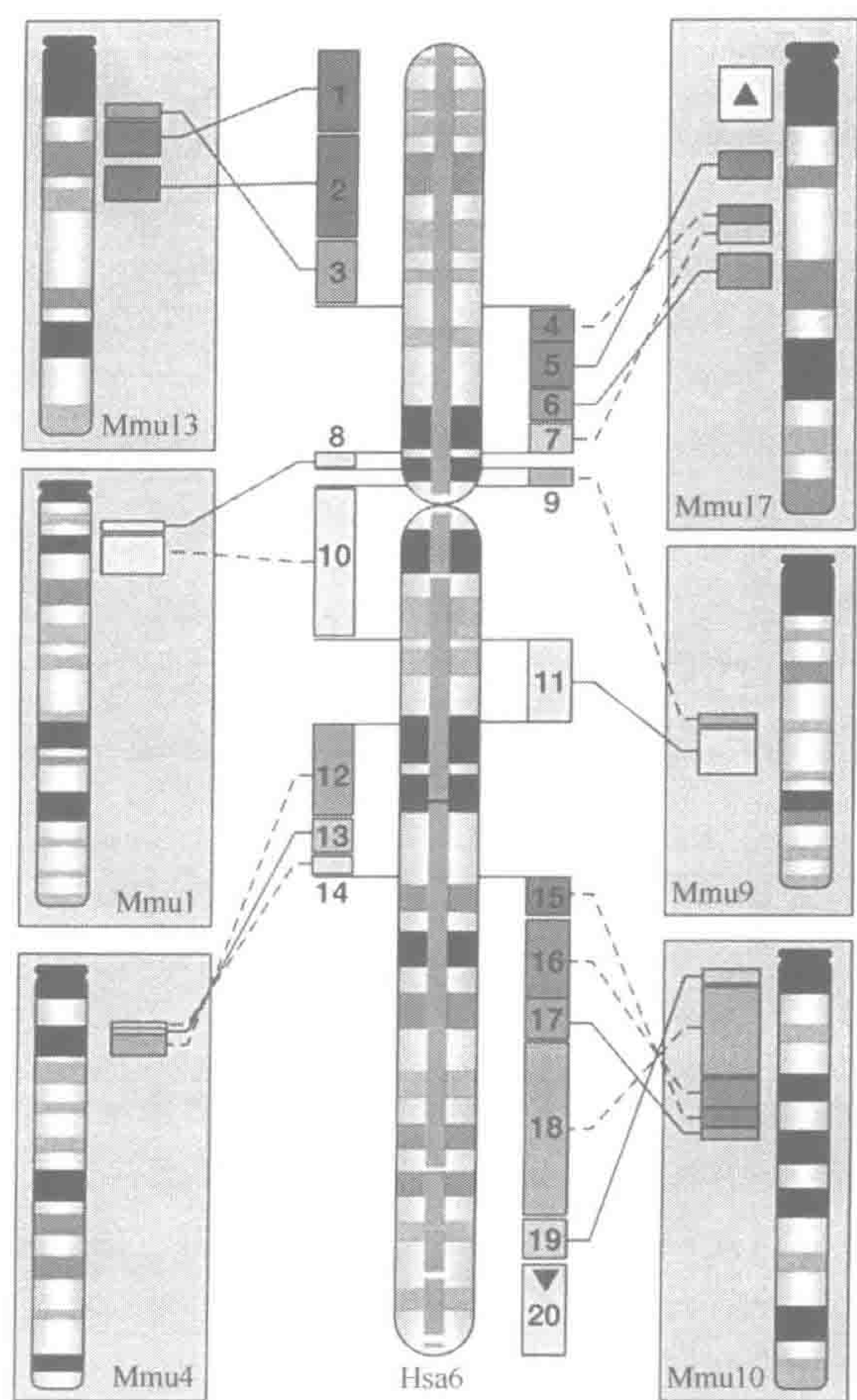


图 14.8　人 6 号染色体（Hsa6）和小鼠染色体（Mmu）之间的保守片段

通过比较人与小鼠的基因组序列，鉴定了 20 个保守同线区域。虚线代表与小鼠染色体图表反向相关的区域。染色体核型模式图中央的蓝色横线代表分析中所包含的重叠群。摘自 Gregory *et al*. (2002). *Nature* 418，743～750. Nature Publishing Group 授权。

框 14.2　小鼠基因作图

有几种简单快速的方法可以用于在小鼠中进行表型或 DNA 克隆的作图。结合转基因小鼠构建技术，小鼠已成为人类特别有价值的对照。方法如下：

种间杂交（*Mus musculus*/*Mus spretus* 或 *Mus castaneus*）

物种在许多多态基因座上具有不同的等位基因，这使得很容易辨认出标志等位基因的来源。主要通过两种方法：

▶ 构建标记框架图谱。一些实验室已经构建出大量系列的 F2 回交小鼠。任何标记或克隆的基因可以快速地排列在由收集的回交小鼠中两个重组体断裂点限定的一个染色体的小片段上。例如，欧洲协作的回交是由 *M. spretus*/*musculus*（C57BL）杂交产生的。500 只 F2 小鼠通过与 *spretus* 回交，500 只通过与 C57BL 回交产生。框架图谱中的所有微卫星标记都标记在每个小鼠中。

▶ 定位一种新的表型。必须建立特异的杂交来完成，但与人类不同，任何数量的 F2 小鼠都可以繁殖用于绘图达到所需的分辨率。*mustulus*×*castaneus* 的杂交比 *musculus*×*spretus* 更易于繁殖。

框 14.2 小鼠基因绘图（续）

重组近交系

这些是通过将杂交后代进行系统的近交而得到的，例如广泛使用的 B×D 株是源自一个 C57BL/6J×DBA/2J 杂交后代近 60 代近交产生的一组 26 个系。它提供了具有固定重组体点的染色体嵌板的无限储备。DNA 可用作公共资源，种系的功能更像 CEPH 家族对人类的作用（节 13.4.2）。重组近交系株特别适用于定位数量性状（节 15.6.8），它可以在每个亲代株中鉴定，并且在每个重组体型的许多动物中取得平均值。与来自物种间杂交的小鼠相比，它在指定区域内可能更难找到可以区分两个来源种株的标记，并且由于数量少，因此分辨率低。

类等基因系

除一个特定基因座外，它们都是相同的。是由重复回交产生的，可以用于判断恒定背景下仅改变单一的遗传因子的效果。

Silver（1995）总结了小鼠的遗传学（见进一步阅读），Copeland 和 Jenkins（1991）描述了种间杂交的应用。

14.4 应用染色体畸变

染色体畸变（chromosome abnormality）有时会替代连锁分析，提供定位致病基因的另一种方法。由于疾病在正常情况下是散发的，像许多严重的显性性状，染色体畸变可能是得到候选基因的唯一方法（节 14.6.1）。运气好的话，它们甚至可以直接指出精确的位置，而不像连锁分析是限定一个候选区。平衡异常（易位或倒位）尤其有意义。机智的医生们在发现此种患者时发挥了关键作用（框 14.3）。亚显微的缺失与隐性易位至少与可见的染色体畸变具有同样的价值。

框 14.3 存在染色体畸变的标志

临床医师们通过发现携带致病染色体畸变的患者而为鉴定致病基因做出主要贡献。

具有细胞遗传学变异的患者有标准的临床表现

假如一种致病基因已经被定位在特定位点上，随后该病的患者被发现携有染色体畸变影响同一位点，则这个染色体畸变最可能引发了此疾病。

- 携有平衡易位和倒位的患者常在致病基因上或其附近存在断裂点。克隆它们的断裂点可以为鉴定致病基因提供最快捷的途径。
- 带有中间缺失的断裂点也许会与致病基因有一段距离，但如果缺失片段比目前的候选区域小，确定断裂点则有助于定位基因。

大部分这样的患者会有新生突变。一些研究者认为，在所有有新发突变的患者中进行染色体分析是值得付出科研努力的。

附加的智力低下

患者可能患有典型的孟德尔遗传病，除此之外还存在严重的智力低下。这可能是偶然的，但这样的病例可能由于致病基因及额外的邻近基因的缺失所致。大的染色体缺失几乎总引发严重的智力低下，这反映了我们的基因在胎儿脑发育中占有很高比例。当患者有新发突变时，有理由进行细胞遗传学和分子分析。

框 14.3　存在染色体畸变的标志（续）

连续基因综合征

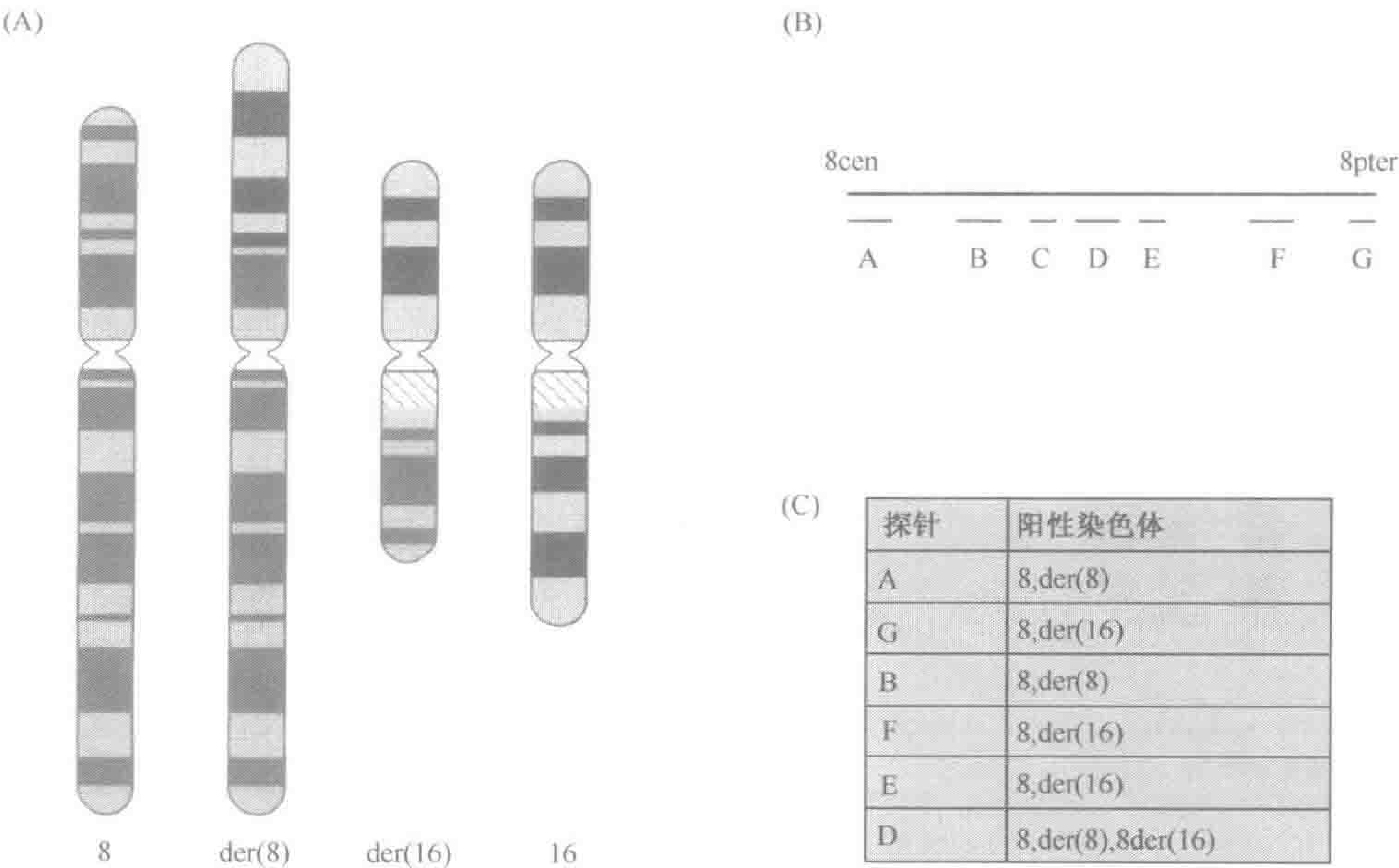
一位患者同时患有几种不同的遗传病很罕见。这可能仅仅因为运气太差，但有时是因为连续系列基因的同时缺失。连续基因综合征已在节 16.8.1 中描述了；对 X 连锁疾病，他们有特别深入的阐释。

14.4.1　具有染色体平衡畸变和无法解释的表型的患者是有意义的

平衡易位或倒位没有增加或减少任何物质，不被认为会对携带者的表型产生影响。假如一个具有明显的染色体平衡畸变的人在表型上是异常的，则有三种可能的解释：

- ▶ 这个发现是偶然的；
- ▶ 重排实际上是不平衡的——有未被发现的物质的缺失或增加；
- ▶ 染色体断裂中的一种引起的疾病。

如果染色体断裂破坏了基因的编码序列或将它与附近的调控区分开，可以产生功能丢失的表型。或者也产生功能获得，例如将两个基因的外显子剪接起来产生一个新的嵌合基因（这在遗传疾病中少见而在肿瘤发生中常见，见 17 章）。在任一情况下，断裂点给致病基因的精确物理位置提供了有价值的线索，用 FSIH 对断裂点的精确定位最容易被确定（图 14.9）。此种方法的一个有力例子是鉴定 Sotos 综合征基因（节 14.6.1）。然而定位线索是不可靠的；有时通过影响大范围染色质结构域断裂点，可以改变位于几百 kb 之遥的基因的表达（框 14.4）。



探针	阳性染色体
A	8,der(8)
G	8,der(16)
B	8,der(8)
F	8,der(16)
E	8,der(16)
D	8,der(8),8der(16)

图 14.9　应用荧光原位杂交确定易位断裂点

(A) 细胞遗传学确定易位 t（8;16）(p22;q21)。(B) 正常 8 号染色体中部分断裂点区域的物理图，显示了 7 个克隆的大致位置。(C) 连续的 FSIH 实验结果。断裂位点在克隆“D”所代表的序列中。正常情况下，这个结果应该用来自第 16 号染色体的克隆来确定。

框 14.4　位置效应——鉴定致病基因的陷阱

通常，看起来基因在染色体上或多或少是随机排列的，并且精确的排列或次序是无关紧要的。然而在果蝇中，众所周知，局部 Mb 大小的染色质组织可以影响其基因表达，特别是如果被放到异染色质内或其附近，基因就沉默了。在小鼠和人类，同样也是真实的。

对转基因表达的研究显示（节 20.2.3），正确的组织特异性基因表达可以依靠位于距离基因编码序列几百 kb 的序列。已知几个人类的实例，易位断裂点影响了直至 Mb 以外的基因的表达，无虹膜（MIM 106210）和 *PAX6* 基因，躯干发育异常（MIM 211970）及 *S0X9* 基因，已在节 10. 5.1 中提及。

因此，平衡易位断裂点不必定位于被其灭活的基因内或其附近，这削弱了它们作为克隆致病基因工具的价值。

即使易位断裂点破坏了一个基因，我们仅丢失了这个基因两个拷贝之一的功能。除非产物水平减少 50%会引发问题（单倍性不足，节 16.4.2），否则不会有表型效应。由于 X 失活，使女性中 X-常染色体易位成为特殊的例子。失活是随机的，但包含失活易位 X 的细胞通常经受致死性的遗传不平衡（图 14.10），因此此种易位的女性携带者将全部由正常 X 失活了的细胞组成。如果易位断裂点破坏了 X 染色体上的基因，则女性患者只保留基因的无功能拷贝。世界上有约二十多名女性患者由于 X-常染色体易位而患有 DMD。每种易位都有不同的常染色体的断裂点，而 X-断裂点总是 Xp21。对这些女性患者的研究为早期将 DMD 基因定位在 Xp21 上的连锁分析工作提供了补充，其中之一带有 X;21 易位，为克隆 DMD 基因提供一种方法（见下文）。

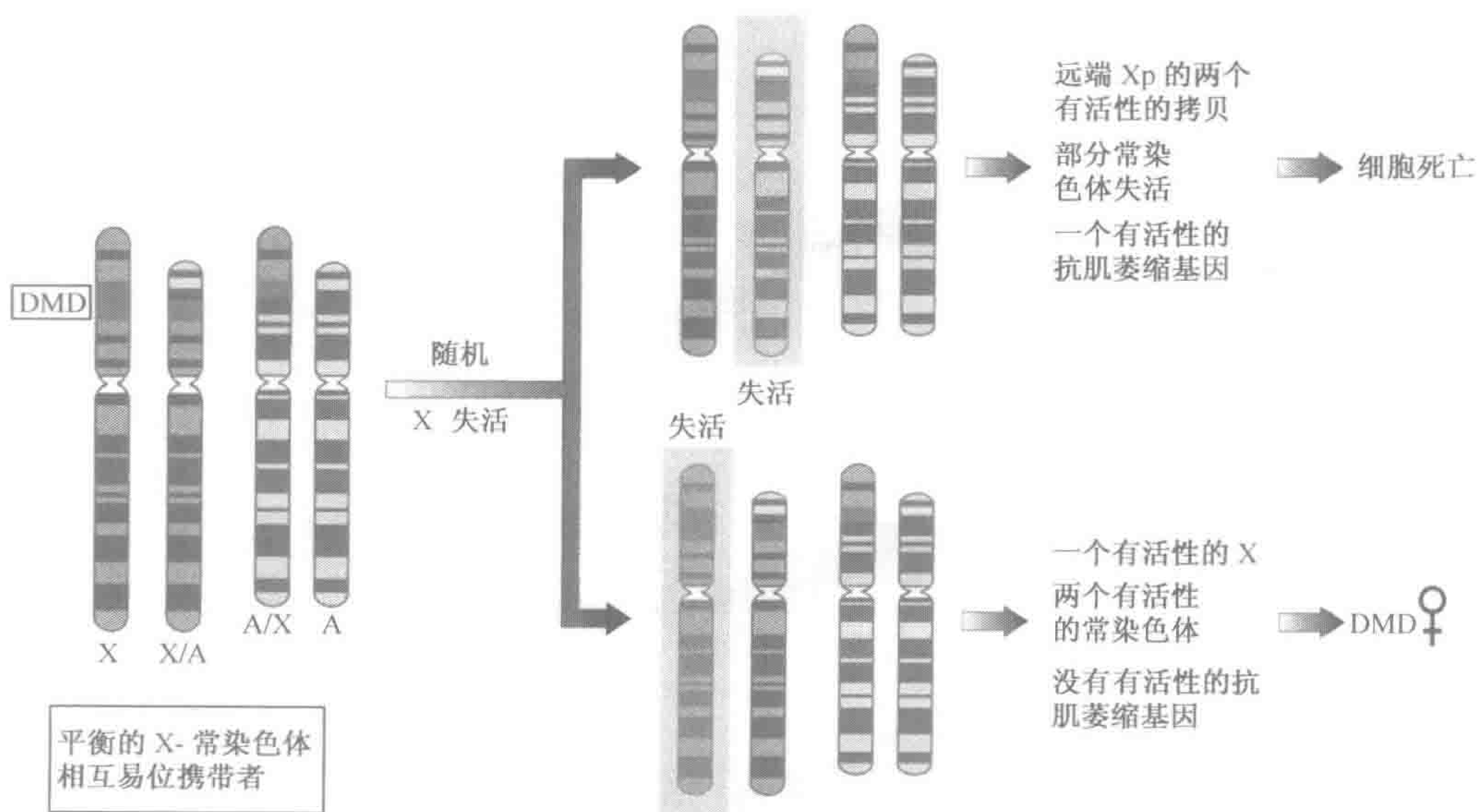


图 14.10　带有 Xp21-常染色体易位的女性 DMD 患者中发生的非随机 X 失活

易位是平衡的，但 X 染色体断裂点破坏了抗肌萎缩基因（红色框）。X 失活是随机的，但由于致死性遗传不平衡，灭活了易位的 X，细胞死亡。胚胎完全从正常 X 失活的细胞发育而来，致使该女性没有功能性抗肌萎缩蛋白基因。结果不能产生任何抗肌萎缩蛋白而导致 DMD。

将探求的基因安放在一已知序列附近的重排为获得未知基因提供了一个直接途径。模式生物的许多基因是通过由转基因或可移动元件插入所导致的突变克隆的。人类对 DMD 基因的尝试就是采用这种方法。一位 DMD 的罕见女性患者有 Xp；21p 的易位。已知 21p 被大批重复的 rRNA 基因所占据（节 9.2.1），Worton 的小组制备了基因组文库并开始寻找包括 rDNA 和 X 染色体序列的克隆。结果分离出 XJ（X-连接）克隆，它被证明位于抗肌萎缩基因的内含子 7 中（Worton and Thompson，1988）。

上述的第二种可能解释同样不能忽视。对具有新的明显的平衡畸变和表型的患者的研究表明，大部分患者实际上具有更复杂的染色体重排，通常包含亚显微的缺失。即使丢失了 1 Mb 的 DNA，在标准的细胞遗传学标本制备中也不会被发现。大于几 kb 的缺失可以被 FSIH（图 14.12）检测到，杂合子易位携带者的更小缺失最好在体细胞杂种分离出衍生染色体（框 8.4）后通过 PCR 进行研究。

14.4.2 具有两种孟德尔疾病或一种孟德尔疾病加上智力低下的患者可能具有染色体缺失

对于鉴定基因，染色体缺失比平衡畸变的价值低一些，因为研究重点是整个缺失区域而不是特异的断裂点。然而，缺失有助于鉴定几种重要的致病基因，包括早期的里程碑似的成就：抗肌萎缩基因的鉴定。

这里的出发点是一个叫“BB”的男孩，他患有 DMD 并具有细胞遗传学上可见的 Xp21 缺失。将技术难度大的**消减克隆**（subtraction cloning）应用于分离与“BB”中缺失的序列相当的正常 DNA 中的克隆（Kunkel *et al*.,1985）。消减文库中的特殊 DNA 克隆随后用作对正常人和 DMD 患者 DNA 样品进行 Southern 印迹杂交的探针，以探测来自正常人群和 DMD 患者的 DNA。一个克隆，pERT87-8，在 7%细胞遗传学正常的 DMD 患者的 DNA 中检测到缺失。并且还检测到多态性，通过家系研究显示，该多态性与 DMD 紧密连锁。这些结果表明，pERT87-8 比以前分离的任何克隆都更靠近 DMD 基因（事实上它在基因之内，在内含子 13 中）。通过染色体步查分离得到其他邻近的基因组探针，然后通过动物印迹寻找保守序列，用于筛查肌肉 cDNA 文库。由于抗肌萎缩蛋白 mRNA 含量低，及我们目前已知的其外显子片段小且位置分散广泛，所以寻找 cDNA 克隆的工作很不容易，但最终此克隆还是被鉴定了，并且随后定性了整个奇异的抗肌萎缩蛋白基因（图 10.14）。

最近应用缺失的例子涉及 *PHEX* 基因，它在 X 连锁显性抗维生素 D 佝偻病（MIM 307800）中发生了突变。该基因定位于 Xp 上的一个很小的区域内，但在 150 名受累男性中的 4 人中发现了亚显微缺失，使注意力集中于候选区的一小部分上（HYP 协作组，1995）。缺失对 X 连锁情况尤为有用，因为受累男性患者没有来自正常染色体的干扰。

微缺失（microdeletion）通常被认定是大量无法解释的遗传综合征的病因。由于缺失区域很小，对于确认病理改变涉及的基因会特别有价值。然而，直至今日，仍没有办法去系统地搜寻它们。如果顺利的话，敏感的高分辨的比较基因组杂交技术（框 14.5）的发展可以补救。用 FSIH（图 14.12）或脉冲电场凝胶电泳及 Southern 印迹确定可疑的缺失是很重要的。PCR 扩增产物失败可能是由于引物结合位点的一个序列变化而不是缺失。

框 14.5　CGH 检测亚显微的染色体不平衡

比较基因组杂交（comparative genomic hybridization，CGH）用于检测部分的单体或三体染色体的缺失或扩增。正如在节 17.3.3 所述，CGH 是基于使测试 DNA 和对照 DNA 竞争与靶标杂交。靶标可以是显微镜载玻片上的正常染色体，应用标准的荧光原位杂交（节 6.3.4）或是排列在载玻片上的一套 BAC 克隆（阵列-CGH）。染色体区域或 BAC 克隆中测试和对照样品的拷贝数的不同以不同着色的荧光信号区域或点呈现出来（图 17.3）。阵列-CGH 可在具有先天异常的患者中进行微缺失或微扩增的全基因组扫描。

14.5　确定候选基因

候选基因必须分别检测，以查看是否有表明它们的突变真能引起正在讨论中的这种疾病的好的证据。证明候选基因可能是可以通过各种方法来进行的疾病基因。

- **突变筛查**　筛查候选基因中患者特异性突变是目前最常用的方法，因为它应用广泛且比较快速。16 章讨论了特异突变出现在特定疾病中的原因，而且在 18 章中描述了检测突变的方法。在几位不相关的患者之中鉴定出突变强烈暗示已选择了正确的候选基因，但正式的证据需要额外的根据。
- **体外正常表现型的修复**　假如已证实患者细胞具有突变的表型，我们可以检测转染了已克隆进入表达载体的候选基因的正常等位基因（节 5.6.1）是否能“挽救”突变并恢复正常的表型。但并不是所有突变表型都是可逆转的，所以阴性结果也没有必要排除候选基因。
- **建立疾病的小鼠模型**（节 20.4.3）　一旦一种推测的致病基因被鉴定，而且不存在相关的突变体，就可以构建转基因小鼠模型。功能丢失的表型可以通过对小鼠种系的基因打靶来制作敲除模型（节 20.4.4）。对于功能获得的表型，疾病等位基因必须被引入到小鼠种系中，突变小鼠希望能与患此病的人类具有某些相似之处，即使正确的基因已鉴定出来，这种预期可能不是总能碰到的。

14.5.1　突变筛查证实一个候选基因

对于大部分患者是携带独立突变的疾病来讲突变筛查应是直截了当的。典型的是这些严重早发的常染色体显性或 X 连锁疾病，该病的表型由基因功能丢失所致。如节 16.3.2 中解释的，如果用节 18.13 所述，一种或多种突变筛查方法检测到正确的人类基因，则来自不相关患者的一批样品常常会显示不同突变的多样性。希望这其中包含一些对基因表达具有明显缺失效应的突变，如无义突变，移码突变等。图 16.1 显示了一个例子。需要检测正常对照以证实任何改变都不是常见的群体变异体。

在其他环境下，鉴定突变并解释突变筛查会更加困难。

- **毋庸置疑的基因座异质性**　通常几种不同的基因突变可以产生几乎相同的表型，因此，一批未被选择的患者样品可能在不同基因上有病理性突变。如果检测的候选基因只引起小部分的病例，则大部分样品将不会呈现该基因的突变。理想化地，应该用于自己证实与候选区连锁的家系的样品，但这可能是行不通的——对于隐性及某

些显性疾病，进行独立的连锁分析，家系规模太小。而且一些严重的显性疾病，患者是没有家族史的散发病例。

- **突变的同质性**　如果大部分不相关的患者具有相同的序列改变，这可能是致病性突变或可能是一个与真正突变强烈连锁不平衡的罕见变异体。需要功能的证据以证明改变是致病性的。
- **突变不是明确地致病性的**　与对基因表达无主要影响的罕见中性变异体对比，可能很难断定错义突变是致病性的。有助于决定一种序列的改变是否为致病性的一些原则已在框 16.4 中指出。
- **突变也许很难找到**　除了筛选一个有许多外显子的大基因的实际问题外，一些突变无法通过基因组 DNA 的 PCR 检测被确认。例如，检测破坏Ⅷ因子基因的大片段倒位（节 11.5.5 和图 11.20），或激活深藏在内含子中隐匿剪接位点的 *CFTR* 3849＋10kb C>T 突变（节 16.4.1），将需要分别应用 Southern 印迹或 RT-PCR 技术。

14.5.2　一旦候选基因被确定，下一步是了解它的功能

鉴定涉及遗传性疾病的基因为若干研究路线打开了道路。鉴定突变的能力会迅速提高诊断和咨询，正如 18 章中所述。了解分子病理（为什么突变基因引起疾病，见 16 章）也可能促进对相关疾病的领悟，而且有望最终得到更有效的治疗。

第二条路线的查询涉及到**基因产物的正常功能**。在 DMD 抗肌萎缩蛋白基因被发现以前，我们对肌肉细胞的收缩系统锚定在肌膜上的方式一无所知。对功能结构域和基序的分析，及在小鼠、果蝇、线虫和酵母中寻找可进行实验性操作的同源体是此项工作的有力工具。这些大的题目分别涵盖在 19 和 20 章；从下面的信息可以看出通过数据搜索能获得信息种类的预测，取自“遗传性听力丧失”首页（http://www.uia.ac.be/dnalab/hhh/），它描述了在一个庞大的哥斯达黎加家系中（OMIM entry 124 900），通过一个常染色体听力丧失基因座（*DFNA1*）的定位克隆而鉴定的基因。

人类 *DFNA1* 蛋白产物 DIAPH1，小鼠 *p140mDia* 及果蝇 *diaphanous* 是酵母蛋白 *bnilp* 的同源体，整个蛋白质都是高度保守的。编码这些蛋白的基因是 *formin* 基因家族的成员，它也包括：小鼠肢体畸形基因、果蝇 *cappuccino*、曲霉 *nidulans* 基因 *sepA* 及裂殖酵母属 *pombe* 基因 *fus1* 和 *cdc12*。这些基因参与细胞质分裂及细胞极性的建立。所有 *formin* 在 N 端区域都有 Rho-结合结构域，在每个序列中心区域有多聚脯氨酸延伸，并且在 C 端区域具有 *formin* 同源结构域。

14.6　以 8 个例子阐明鉴定致病基因的各种方法

14.6.1　通过染色体畸变直接鉴定致病基因：Sotos 综合征

Sotos 综合征（MIM 117 550）以过生长、畸形面容和智力低下为特征。大多数病例是散发的；受累个体很少生殖，并且没有完好的家系用于连锁绘图。一位 Sotos 患者被发现具有平衡易位 46，XX，t(5;8)（q35;q24.1）。通过 FISH（图 14.9）鉴定了跨越断裂点的 PAC/BAC 及黏粒克隆。对断裂点周围进行测序，揭示了与小鼠的 *NSD1* 基因同源的部分基因组序列。人类 *NSD1* 基因被克隆并鉴定，显示其被易位（图

14.11）所破坏。那可能仅仅是偶然的。*NSD1* 是 Sotos 综合征中突变基因的证据，38 个孤立 Sotos 患者中 4 位被证明有点突变，30 位（Kurotaki *et al*.,2002）中的 20 人发现了微缺失（图 14.12）。需要注意的是，即使基因在物理结构上没被破坏，但它可能被染色体重排所影响（框 14.4），所以不是总能如此直截了当地发现，易位是引发疾病的原因。

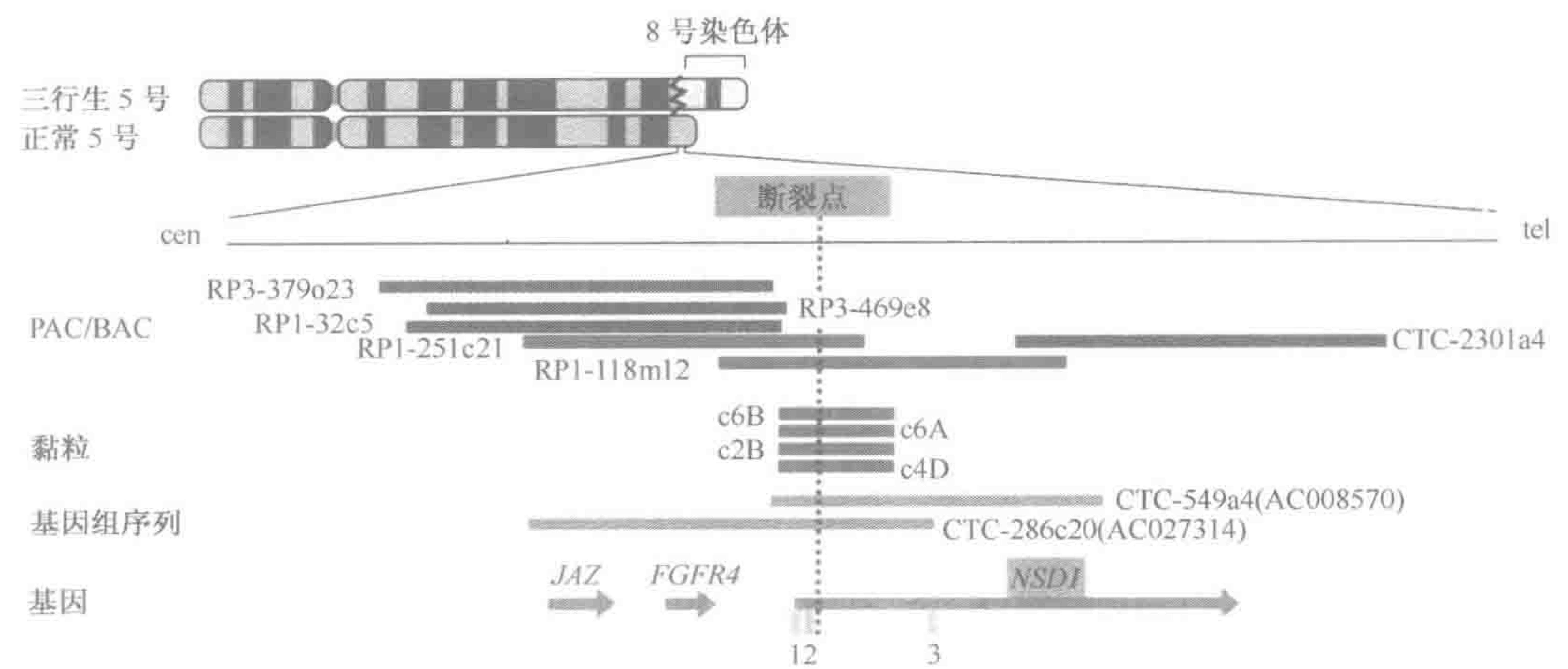


图 14.11　一平衡的 5;8 易位破坏了 Sotos 综合征患者的 *NSD1* 基因

摘自 Kurotaki *et al*.（2002）Nat. Genet. 30，365～366. Nature Publishing Group 授权。

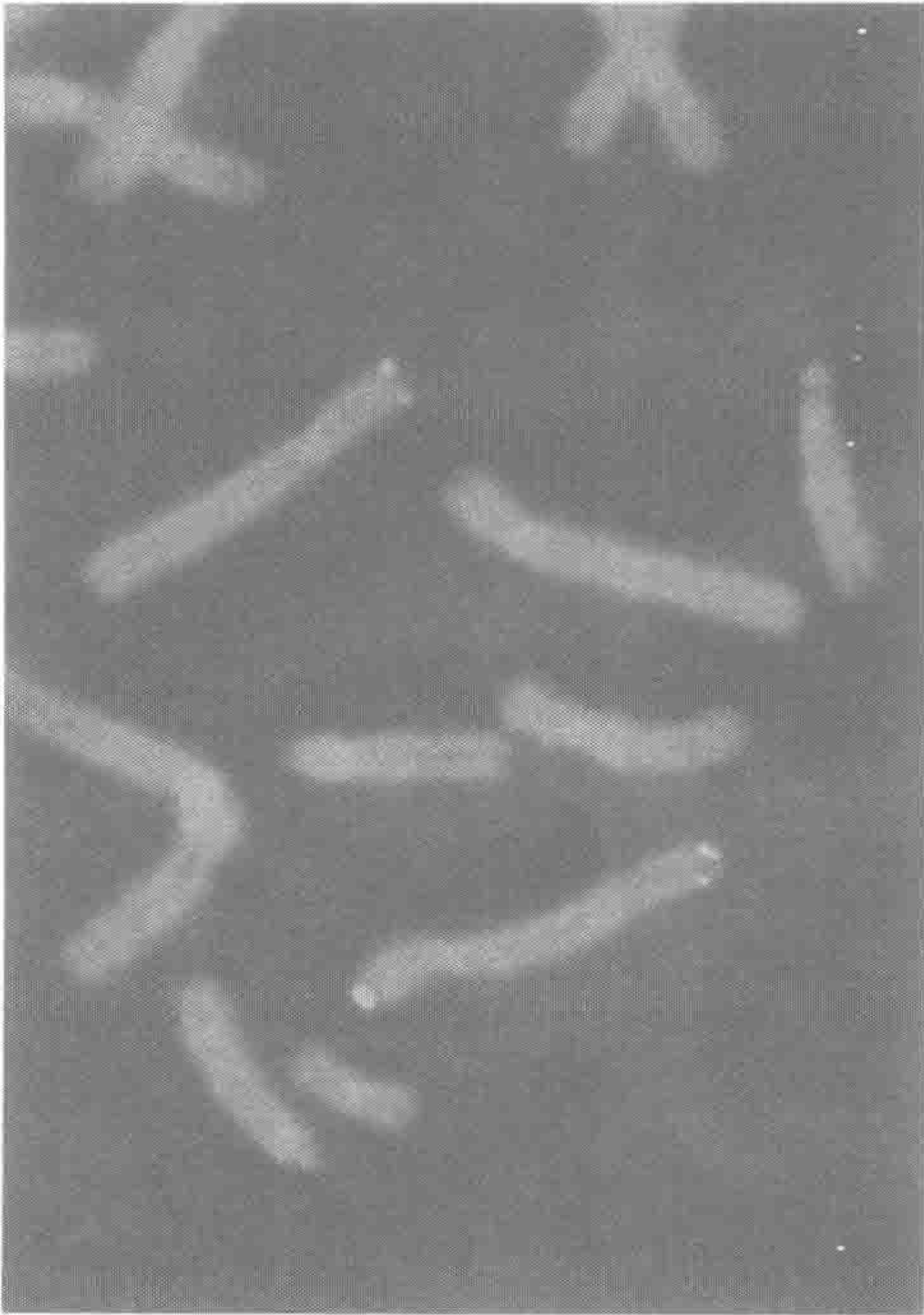


图 14.12　荧光原位杂交证实 Sotos 综合征患者的 *NSD1* 微缺失

通过识别 5pter 上序列的红色 FSIH 探针鉴定了 5 号染色体的两个同源体。绿色 FSIH 探针是源自 5 号长臂末端含有 *NSD1* 基因的一个 BAC；此序列在 5 号染色体的一个拷贝上发生缺失。摘自 Kurotaki *et al*.，（2002）Nat. Genet. 30，365～366. Nature Publishing Group 授权。

14.6.2　纯粹的转录物作图：Treacher Collins 综合征

Treacher Collins 综合征（TCS；MIM 154500）是颅面发育的一种常染色体显性病，具有包括外耳和中耳的异常、下颌骨发育不全、颧骨复合及腭裂在内的多种表型。连锁分析最初建立于 1991 年，与 5q31-q34 的标记连锁。因为当时此区域内的标记无法提供足够的信息，新的微卫星被分离出来用于缩小候选区至 5q32-33.1。构建了跨越该区域的组合遗传的辐射杂交图谱，并且直至 1994 年，研究小组组合了 YAC 重叠群。它被转化为黏粒重叠群，用 cDNA 文库筛查和外显子捕获产生转录图。在关键的区域内发现了至少 7 个基因。新一轮的标记分离和交错分析产生了重叠重组体的混乱图像，但最终还是从胎盘文库中分离出候选的不完整的 cDNA。Northern 印迹和动物印迹显示，基因广泛表达并在物种间保守，搜寻显示没有显著的同源性。外显子-内含子结构被确定，突变分析证实了在 5 名不相关的患者中存在不同的突变。

TCOF1 基因的分离（Treacher Collins 综合征协作组，1996）以其完全形式证明了位置克隆。没有发现相关的染色体畸变（有 4 名 TCS 患者有染色体易位或缺失，但来自每个断裂点的标记在家族研究中显示没有连锁，所以推测这些病例都是偶发的）。没有连锁不平衡——由于 60%的病例是新发突变，因此这并不奇怪。此候选区域富含基因，因此有许多可能的候选基因，而且最终被鉴定的基因不具有使之成为特别可能的候选基因的特征。基因产物是核仁的磷酸蛋白，但为什么突变会确切地引起 TCS 仍未知晓。

14.6.3　大规模测序和寻找同源体：branchio-oto-renal 综合征

常染色体显性 branchio-oto-renal 综合征（BOR；MIM：113650：鳃状瘘管，外耳和内耳的畸形伴听力丧失；肾脏发育不良或缺失）以一个受累患者具有重排的 8 号染色体为线索被定位于 8q13。通过进一步作图以及在提及的患者中发现的亚显微染色体缺失将最初的 7cM 间距限定至 470～650kb。通过应用候选区内或其附近的标记筛选基因组文库，P1 和 PAC 克隆被分离出来，而且重叠群的空隙由染色体步查填充。跨越候选区的最小的覆瓦式途径，共发现 3 个 P1 和 3 个 PAC 克隆。

鉴定重叠群中的基因决定进行大规模测序。在 EMBL 和 GenBank 蛋白质与核酸数据库中检测序列，发现了已获得序列的一部分与果蝇发育基因无眼（*eya*）之间的同源性。在基因组序列中寻找可读框，它被翻译并与 *eya* 氨基酸序列进行比较。结果，鉴定的 7 个预测的外显子在氨基酸水平呈现出与预测的 *eya* 蛋白具有 69%的一致性和 88%的相似性。接着，从 9 周的胎儿总 mRNA 文库中分离出人类 cDNA，并且在 42 名不相关的 BOR 患者中证实，存在 7 个突变，命名为 *EYA1*（Abdelhak *et al.*,1997）。

伴随着基因组序列数据库的日趋完善，在此应用的各种分析成为越来越标准的方法。此时，远缘相关有机体间基因的同源性在氨基酸序列中比在 DNA 序列或表型中可能更显而易见。在这个阶段，基因产物的功能仍未鉴定，也不清楚为什么果蝇的表型包括复眼的减少或缺失，而人类则没有眼疾。正如经常伴随定位克隆那样，鉴定致病基因只是了解此种综合征的开始。

14.6.4　通过功能限定位置的候选基因：视紫质和纤维蛋白原

人类感光视紫质（*RHO*）基因于 1984 年被克隆，1986 年被定位于 3q21-qter。在

疾病中涉及包括遗传性视黄醛退化在内的视网膜色素异常（RP）的各种形式，它因视网膜色素凝结而表现为显著的进展性视力丧失为特征。尽管 *RHO* 可能是某些形式的 RP 的候选基因，但它只是编码已知的参与光信号转导蛋白质的众多基因之一。然而在 1989 年，在一个较大的爱尔兰 RP 家系进行的连锁分析将他们的致病基因定位至 *RHO* 的 3q 附近。现在这是一个重要的候选基因，患者特异性突变在一年之内被鉴定出来（见 OMIM 登录号 180380）。

Marfan 综合征的表型（MFS. MIM 154700：长骨的过度生长；关节松弛；晶状体脱位；易患主动脉瘤）提示在结缔组织中存在异常。连锁分析将 MFS 基因定位于 15q，当用原位杂交将结缔组织蛋白纤维蛋白原基因定位于 15q21.1 后，它成为一个明显的位置候选基因。很快在 MFS 患者中证实了突变［Mckusick（1991）对研究背景进行了讨论］。

14.6.5　通过比较人类与小鼠的图谱鉴定位置候选基因：*PAX3* 和 Waardenburg 综合征

1 型 Waardenburg 综合征（WS1，MIM 193500）体现了人类-小鼠比较的价值。图 4.5C 显示了这种常染色体显性的系谱。WS1 特征性的色素异常和听力丧失是由于受累部分的黑色素细胞缺乏而导致的，包括内耳，正常听力发育过程中耳蜗血管纹的形成需要此处的黑色素细胞。通过对一个受累患者染色体畸变的描述帮助了连锁分析，最终将 WS1 基因定位于 2q 的远端。此染色体区域与部分小鼠 1 号染色体呈现强烈的保守同线性。从这点看来，可能的小鼠同源体出现了。*Splotch*（*sp*）小鼠突变体具有黑色素细胞斑片状缺失引起的色素异常。这可能是胚胎神经嵴缺陷的结果。虽然两种表型之间有各种差别，但看起来很可能 *Sp* 与 WS1 是由种内同源性基因突变引起的。

两个基因都没被鉴定，但当小鼠 *Pax-3* 基因被定位于 *Sp* 基因座附近时，出现了位置候选基因。*Pax-3* 编码一种转录因子，它在包括神经嵴在内的神经系统发育过程中的小鼠胚胎中表达。小鼠 *Pax-3* 的序列与先前报道的未定位的人类基因组克隆 HuP2 的限制性序列几乎一致。这种发现促进了对 *Pax-3* 和 HuP2 的突变筛查，并导致在 *Splotch* 小鼠和 WS1 患者中鉴定了突变（Strachan and Read 综述，1994）。因为从根本上说这个基因显然是种间同源体，所以 *HuP2* 基因被重新命名为 *PAX3*。

14.6.6　从体外功能推测：Fanconi 贫血

Fanconi 贫血是一种隐性疾病，具有多种先天异常，特别是桡骨发育不良，易患骨髓衰竭和恶性肿瘤，尤见急性骨髓粒细胞性白血病。根本问题是 DNA 损伤修复功能的缺陷。此种缺陷在细胞培养中可以观察到对如双环氧丁烷之类的 DNA 交联剂高敏感性。细胞融合实验将 Fanconi 患者分成至少 8 个互补组（A-H）：当融合时，来自不同组患者的细胞互相弥补，然而那些来自同一组患者的细胞仍保持缺陷。通过检测 cDNA 文库克隆修复来自 Fanconi 患者 C 组细胞的对双环氧丁烷敏感的能力，Strathdee 等（1992）分离出 Fanconi 贫血 C 组基因（*FANCC*；MIM 227645）基因。后来应用相似的方法，鉴定了 A 组（*FANCA*）和 G 组（*FANCG*）基因。功能克隆的另一个用途是用转移的染色体或克隆矫正肿瘤细胞系的生长失控，已用于辅助定位和随后对肿瘤抑制基因的鉴定（节 17.4）。

14.6.7　从体内功能推测：肌球蛋白 15 和 DFNB3 耳聋

有时，通过候选区内野生型 DNA 转基因所挽救的突变表型，可以鉴定一个小鼠致病基因。这种策略首先应用于鉴定一种时钟基因（Antoch *et al*.,1997），最近成为鉴定人类 *DFNB3* 耳聋基因（Probst *et al*.,1998；图 14.13）的关键步骤。比较作图显示，*DFNB3* 在人类定位于一个位置，与小鼠中 *shaker-2* 耳聋基因的位置相对应。用来自 *shaker-2* 候选区的野生型 BAC 构建 *shaker-2* 转基因鼠，鉴定了一个矫正了表型的 BAC，并且被证明为包含有一种不寻常的肌球蛋白基因 *myo15*。人类 *MYO15* 基因基于与小鼠基因的密切同源性而随后被分离，已确定它定位于 *DFNB3* 候选区内，随后证实，突变存在于 *DFNB3* 受累人群中。

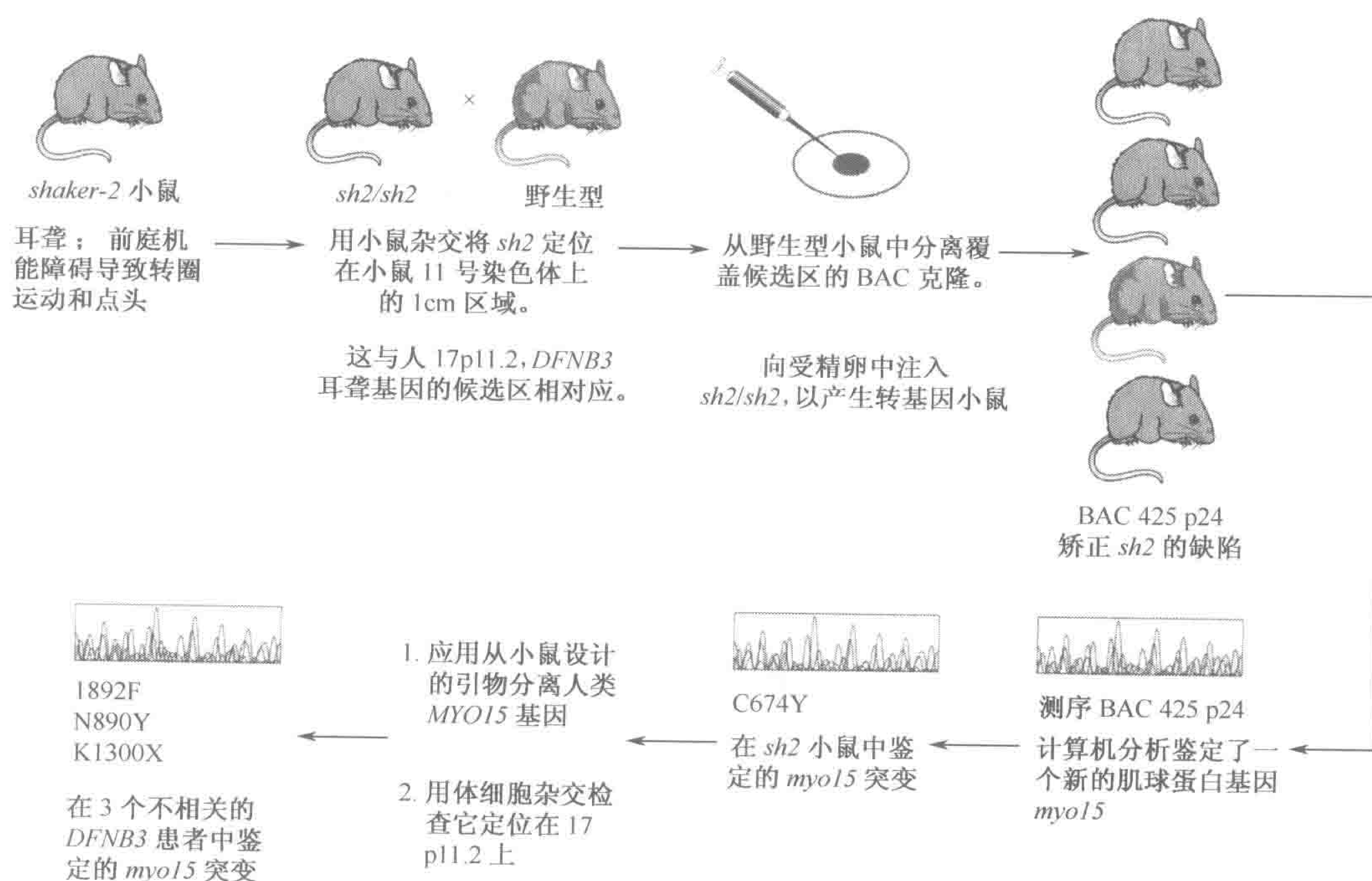


图 14.13　转基因小鼠的功能互补成为鉴定人类致病基因的工具

shaker-2 鼠突变是通过发现能矫正缺陷的野生型克隆而被鉴定的。定位于相应染色体区域、具有相似表型的人类家系证明，在种间同源的基因中有突变。

14.6.8　从表达模式推测：otoferlin

在小鼠内耳中特异性表达的基因 cDNA 文库已成为鉴定耳聋候选基因的有用工具。该文库是通过内耳 cDNA 文库与一种或多种非特异性文库进行消减杂交产生的，以试图排除在各种组织中表达的基因，这是一种高难的技术（Swaroop *et al*.,1991）。我们已经见到连锁分析定位 *DFNB9* 非综合征的耳聋基因（图 13.8）。在接下来的定位克隆（Yasunaga *et al*.,1999）中，来自重要关键区域的一部分基因序列与来自小鼠内耳文库中的一个克隆呈现出 90%的氨基酸一致性和 97%的相似性。此基因命名为 otoferlin，已被完全定性，并在突变筛查时发现该家系中存在一个无义突变。

另一个策略是鉴定小鼠克隆的人类同源体，并且通过 FISH 将其定位（节 2.4.2），它们随后将会成为定位在相应区域内任一耳聋基因的位置候选基因。

（李婷婷　译）

进一步阅读

Silver LM (1995) *Mouse Genetics: Concepts and Applications.* Oxford University Press, Oxford.

Wolfsberg T, Wetterstrand K, Guyer M, Collins F, Baxevanis A (2002) A user's guide to the human genome. *Nature Genet.* **32(suppl).**

参考文献

Abdelhak S, Kalatzis V, Heilig R *et al.* (1997). A human homologue of the *Drosophila eyes absent* gene underlies Branchio-oto-renal (BOR) syndrome and identifies a novel gene family. *Nature Genet.* **15**, 157–164.

Antoch MP, Song E-J, Chang A-M *et al.* (1997) Functional identification of the mouse circadian Clock gene by transgenic BAC rescue. *Cell* **89**, 655–667.

Bartlett SE (2001) Identifying novel proteins in nervous tissue using microsequencing techniques. *Methods Mol. Biol.* **169**, 43–50.

Brown SD, Balling R (2001) Systematic approaches to mouse mutagenesis. *Curr. Opin. Genet. Dev.* **11**, 268–273.

Carter SA, Bryce SD, Munro CS *et al.* (1994) Linkage analyses in British pedigrees suggest a single locus for Darier disease and narrow the location to the interval between *D12S105* and *D12S129 Genomics* **24**, 378–382.

Copeland N, Jenkins NA (1991) Development and applications of a molecular genetic map of the mouse genome. *Trends Genet.* **7**, 113–118.

DeBry RW, Seldin MF (1996) Human/mouse homology relationships. *Genomics*, **33**, 337–351.

Gregory SG, Sekhon M, Schein J *et al.* (2002) A physical map of the mouse genome. *Nature* **418**, 743–750.

Hughes A, Newton VE, Liu XZ, Read AP (1994) A gene for Waardenburg syndrome Type 2 maps close to the human homologue of the *microphthalmia* gene at chromosome 3p12-p14.1. *Nature Genet.* **7**, 509–512.

Hyp Consortium (1995) A gene (PEX) with homologies to endopeptidases is mutated in patients with X-linked hypophosphatemic rickets. *Nature Genet.* **11**, 130–136.

Justice MJ (2000) Capitalizing on large-scale mouse mutagenesis screens. *Nature Rev. Genet.* **1**, 109–115.

Koob MD Benzow KA, Bird TD, Day JW, Moseley ML, Ranum LP (1998) Rapid cloning of expanded trinucleotide repeat sequences from genomic DNA. *Nature Genet.* **18**, 72–75.

Koob MD, Moseley ML, Schut LJ *et al.* (1999) Untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature Genet.* **21**, 379–384.

Kunkel LM, Monaco AP, Middlesworth W, Ochs HD, Latt SA (1985) Specific cloning of DNA fragments absent from the DNA of a male patient with an X chromosome deletion. *Proc. Natl Acad. Sci. USA* **82**, 4778–4782.

Kurotaki N, Imaizumi K, Harada N *et al.* (2002) Haploinsufficiency of *NSD1* causes Sotos syndrome. *Nature Genet.* **30**, 365–366.

Mann M, Hendrickson RC, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473.

McKusick VA (1991) The defect in Marfan syndrome. *Nature* **352**, 279–281.

Pingault V, Bondurand N, Kuhlbrodt K *et al.* (1998) *SOX10* mutations in patients with Waardenburg-Hirschsprung disease. *Nature Genet.* **18**, 171–173.

Poustka A, Pohl TM, Barlow DP, Frischauf AM, Lehrach H (1987) Construction and use of human chromosome jumping libraries from *NotI*-digested DNA. *Nature* **325**, 353–355.

Probst FJ, Fridell RA, Raphael Y *et al.* (1998) Correction of deafness in *shaker-2* mice by an unconventional myosin in a BAC transgene. *Science* **280**, 1444–1447. See also the accompanying paper, **Wang A, Liang Y, Fridell RA *et al.*** (1998) Association of unconventional myosin *MYO15* mutations with human nonsyndromic deafness *DFNB3*. *Science* **280**, 1447–1451.

Putnam EA, Zhang H, Ramirez F, Milewicz DM (1995) Fibrillin-2 (*FBN2*) mutations result in the Marfan-like disorder, congenital contractural arachnodactyly. *Nature Genet.* **11**, 456–458.

Reymond A, Camargo AA, Deutsch S *et al.* (2002) Nineteen additional unpredicted transcripts from human chromosome 21. *Genomics* **79**, 824–832.

Rincón-Limas DE, Lu C-H, Canal I *et al.* (1999) Conservation of the expression and function of *apterous* orthologs in *Drosophila* and mammals. *Proc. Natl Acad. Sci. USA* **96**, 2165–2170.

Robson KJH, Chandra T, MacGillivray RTA, Woo SLC (1982) Polysome immunoprecipitation of phenylalanine hydroxylase mRNA from rat liver and cloning of its cDNA. *Proc. Natl Acad. Sci. USA* **79**, 4701–4705.

Rommens JM, Januzzi MC, Kerem B-S *et al.* (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–1065.

Royer-Pokora B, Kunkel LM, Monaco AP *et al.* (1985) Cloning the gene for an inherited human disorder – chronic granulomatous disease – on the basis of its chromosomal location. *Nature* **322**, 32–38.

Schalling M, Hudson TJ, Buetow KH, Housman DE (1993) Direct detection of novel expanded trinucleotide repeats in the human genome. *Nature Genet.* **4**, 135–139.

Steinmetz LM, Scharfe C, Deutschbauer AM *et al.* (2002) Systematic screen for human disease genes in yeast. *Nature Genet.* **31**, 400–404.

Strachan T, Read AP (1994) PAX genes. *Curr. Opin. Genet. Dev.* **4**, 427–438.

Strathdee CA, Gavish H, Shannan WR, Buchwald M (1992) Cloning of cDNAs for Fanconi's anemia by functional complementation. *Nature* **356**, 763–767.

Swaroop A, Xu J, Agarwal N, Weissman SM (1991). A simple and efficient cDNA library subtraction procedure: isolation of human retina-specific cDNA clones. *Nucl. Acids Res.* **19**, 1954.

Treacher Collins Syndrome Collaborative Group (1996) Positional cloning of a gene involved in the pathogenesis of Treacher Collins syndrome. *Nature Genet.* **12**, 130–136.

Worton RG, Thompson MW (1988) Genetics of Duchenne muscular dystrophy. *Annu. Rev. Genet.* **22**, 601–629.

Yasunaga S, Grati M, Cohen-Salmon M *et al.* (1999). A mutation in *OTOF*, encoding otoferlin, a FER-1-like protein, causes *DFNB9*, a nonsyndromic form of deafness. *Nature Genet.* **21**, 363–369.

第15章　复杂疾病易感基因的定位与鉴定

本章内容

在全世界，遗传学对发病率和死亡率的主要贡献是常见病的遗传因素研究，因此，鉴定疾病的相关基因是医学研究的核心任务。复杂疾病研究的一般策略是：

- 进行**家系、双生子或领养子研究**（family，twin or adoption study）以证明易感性至少部分是遗传的；
- 应用**分离分析**（segregation analysis）来估计易感等位基因的类型及频率；
- 通过**连锁分析**（linkage analysis）定位易感基因座，通常用受累同胞；
- 通过**群体关联**（population association）研究来缩小候选基因范围；
- 鉴定导致易感性的DNA序列变异，确定其生化作用。

下面我们逐一展开这些步骤，通过对8种特定疾病的探讨阐述实际应用情况。最后我们将讨论一个尚无定论的问题，即鉴定易感因素的整体战略是否是医学遗传学未来的关键，是否像有人认为的那样一般是不可能成功的。

15.1 确定非孟德尔遗传性状是否是遗传性的：家系、双生子及领养子研究的作用

15.1.1 λ 值是家族聚集性的衡量指标

毫无疑问，基因决定了孟德尔系谱方式遗传或染色体畸变的相关性状，然而，对于非孟德尔遗传性状，无论是**连续的（数量的）**还是**非连续的（双歧的）**，都有必要证实遗传的决定性作用。为此，明显的途径是证明性状在家系中传递。某种疾病的家族聚集程度可以用 λ_R 的量来表示，λ_R 为受累先证者亲属患病风险与群体患病风险的比值。每类亲属的 λ 值可以计算出来，例如，λ_s 代表同胞的相对风险。λ_R 的数学理论是由 Risch (1990a) 推论出来的。表 15.1 展示了许多精神分裂症研究的总体数据，通过升高的 λ 值证实了家族聚集性，并且正如预期的那样，远亲的 λ 值回落近于 1。

表 15.1　精神分裂症患者的亲属患精神分裂症的风险度：几项研究的总体结果

亲属	患病风险的人数[a]	风险度/%	λ[b]
父母	8020	5.6	7
同胞	9920.7	10.1	12.6
同胞，双亲之一受累	623.5	16.7	20.8
子女	1577.3	12.8	16
子女，双亲均受累	134	46.3	58
半同胞	499.5	4.2	5.2
姑叔姨舅，侄甥	6386.5	2.8	3.5
孙子女	739.5	3.7	4.6
堂表兄弟姐妹	1600.5	2.4	3

a，有精神分裂症风险的人数校正在低于或处于发病年龄段(如 15～35 岁)。

b，λ 值是按照群体患病率设定为 0.8%计算的。

数据由 McGuffin(1984)收集。

15.1.2 共同家庭环境的重要性

遗传学家必须记住，父母给予子女的不仅是环境，还有基因。许多性状在家族中延续是因为有共同的家庭环境——例如，一个人的母语是英语还是汉语。因此，要始终置疑共同环境是否可能是某一家族性状的原因，这一点对于像 IQ 或精神分裂症这样的行为特征更为重要，因为这些行为至少部分地取决于教养。甚至更不能忽视生理性状或出生缺陷：一个家庭可能共用某种不寻常的饮食或某些传统药物而导致发育缺陷。除了家族倾向性外，还需要更多证据来证明一种非孟德尔遗传性状受遗传控制。这些观点在医学文献中应该清楚阐述，却总难以阐明。表 15.5 显示了忽视共同家庭环境所可能发生的情况。

15.1.3 双生子研究受到许多限制

Francis Galton 为数量遗传学奠定了坚实基础，他指出了双生子对人类遗传学的价

值。同卵（MZ）双生子是遗传上等同的两个克隆，而且任何遗传性状将必然是**一致的**（concordant）（两者相同）。无论是遗传模式还是参与基因的数量都如此；唯一例外的是合子形成后的体细胞发生遗传改变所导致的性状（女性 X 染色体的失活方式，所有功能性免疫球蛋白及 T 细胞受体基因等）。异卵（DZ）双生子平均有一半的基因是共同的，就像一对同胞。因此，遗传性状在 MZ 双生子中应该比 DZ 双生子显示出更高的一致性，许多性状确实如此（表 15.2）。

表 15.2　精神分裂症的双生子研究

研究	MZ 一致性	DZ 一致性
Kringlen，1960	14/55 (21/55)	4%～10%
Fischer，1969	5/21(10/21)	10%～19%
Tienari，1975	3/20(5/16)	3/42
Farmer，1987	6/16(10/20)	1/21(4/31)
Onstad，1991	8/24	1/28

表中数字显示的是配对的一致性，亦即通过受累先证者确定的一致性的对数(+/+)和不一致的对数(+/−)。括号中的数据是依对患病广义定义得到的，包括了交界区和表型的患者。一致性也可以通过先证者方法来计算，如果一对双生子都是先证者，计数这一对两次。这样会得出更高值的 MZ 一致性。先证者方法计算一致性比衡量家族聚集性的其他方法更有可比性。只有 Onstad 和 Farmer 使用了目前标准的诊断标准，DSM-Ⅲ。参考文献请见 Onstad 等(1991)和 Fischer 等(1969)。

然而，MZ 比 DZ 双生子一致性高并不能证明一种遗传效应。首先，有一半的 DZ 双生子性别不同，而所有的 MZ 双生子性别相同。即使比较限制在相同性别的 DZ 双生子（正如表 15.2 所示的研究），至少仍存在对 MZ 双生子的行为性状更相似的争议，因为有相同的穿着和待遇，进而比 DZ 双生子共享更多的生活环境。

MZ 双生子出生时分开且分别生长在完全不同的环境可以提供理想的实验（Francis Crick 曾提出一个不切实的建议，即出生的每一对双生子中的一个应该为这一目的而献身于科学）。过去这种双生子的分开比想像的更多见，因为双生子的出生对于一个负担过重的母亲来说有时候是重大打击。很多吸引人的电视节目是关于双生子分离 40 年后重逢，发现双生子有相似的工作，穿相似的衣服，喜欢相同的音乐。然而，分离的双生子作为研究对象也有弊端：

- 任何研究需要以少数有争议的特殊人群为基础；
- 这种分离并非完全的——他们常在出生一段时间后才被分开，并由亲属抚养长大；
- 存在调查偏倚——每个人只想知道分离双生子的惊人相似，而分离双生子的截然不同就没有报道价值；
- 即使在理论上，分离双生子研究也不能区分子宫内环境因素与遗传因素。这对于性倾向选择研究（“同性恋基因”）是很重要的，有人已提出，母体激素会影响宫内胎儿，从而干扰其将来的性倾向选择。

因此，从分离双生子的所有性状的特性来看，它们对人类遗传学研究的贡献相对较少。

15.1.4　领养子研究：鉴别遗传因素和环境因素的金标准

如果分离双生子研究不能区分遗传与家庭环境因素，领养子研究会带来希望。可能有两种研究设计：

- 寻找已知家族遗传某特定疾病的被领养儿，询问该疾病是否在生物学家庭或领养家庭中传递；
- 寻找被领养儿的受累双亲，询问被领养儿是否免于罹患疾病。

Rosenthal 和 Kety 用第一种设计进行了一项著名的（且有争议的）研究（进一步阅读），检测精神分裂症的遗传因素。这项研究的诊断标准曾遭到质疑；提出（争议）并非所有的诊断为真正双盲法。然而，采用 DSM-Ⅲ诊断标准进行的一项独立的重复分析（Kendler *et al.*,1994）得出了大致相同的结论。表 15.3 显示这项研究的后来扩展的结果（Kety *et al.*,1994）。

表 15.3　一项精神分裂症领养研究

	生物学亲属的精神分裂症病例	收养亲属的精神分裂症病例
指标病例(47 个慢性精神分裂症领养儿)	44/279 (15.8%)	2/111 (1.8%)
对照领养儿(年龄，性别，收养家庭的社会地位及领养年数等方面匹配)	5/234 (2.1%)	2/117 (1.7%)

本研究涉及丹麦 14 427 个 20～40 岁年龄段的领养子，其中 47 个被诊断为慢性精神分裂症。这 47 个患者与 47 个同批的对照领养子非精神分裂症配对。数据来自 Kety 等(1994)。

领养研究的首要问题是缺少生物学家庭信息，由于不希望被询问而常导致信息更缺乏。有效的收养登记只在少数国家实行。其次的问题是选择性安置，收养机构常从孩子的利益出发，选择与生物学家庭相似的家庭来安置。虽然领养子研究无疑是检测一种性状遗传程度的金标准，但是因为存在一定难度而主要只在精神病研究上进行，而亲生-养育的争议也特别激烈。

15.2　分离分析用于单纯孟德尔遗传性状和单纯多基因范畴之间性状的研究

正如我们在图 4.6 中所见，单纯孟德尔遗传性状和单纯多基因性状代表一个连续区域的两端。二者之间是少数主要易感基因座控制的**寡基因性状**（oligogenic trait），这些性状的表现可能不同于多基因遗传，也可能更易受主要环境因素的影响。**分离分析**（segregation analysis）是分析性状遗传的主要统计学工具。该方法可以提供是否是易感基因座的证据，至少可以部分地确定其性质。所得结果有助于指导将来的连锁或关联研究。

15.2.1　调查偏倚常是家系数据的一个问题：以常染色体隐性疾病为例

对疾病的研究依赖于病例和家系的收集，所以第一步要考虑原始数据可能存在的偏倚。分离分析需要大量的数据，对数据收集过程的很小偏倚非常敏感。以孟德尔遗传为

例来阐述。假如我们想要证明某疾病是常染色体隐性遗传的，我们会收集一组家系并检测**分离率**（segregation ratio）（受累子女的比例）是 1/4。乍看上去，只要该病不太罕见，似乎就容易做到。但实际上我们样本的预期受累子女的比例并不是 1/4，问题就在于**调查偏倚**（bias of ascertainment）。

假如没有独立的方法来识别携带者，那么只能通过某患儿来鉴定家系。这样，图 15.1 中未画阴影的家系就不能确定，在收集两个孩子的家系中所观察到的分离率是 8/14而不是 1/4。用同样方法确认三个孩子的家系将得出不同的分离率，48/111。任何特定家系的分离率大小都可以从**截短的二项式分布**（truncated binomial distribution）来估计，即（1/4＋3/4)^n 的二项式展开，最后一项（无受累子女）被省略，用于校正实验数据偏倚的最简单方法是框 15.1 所示的 Li 和 Mantel 的方法。

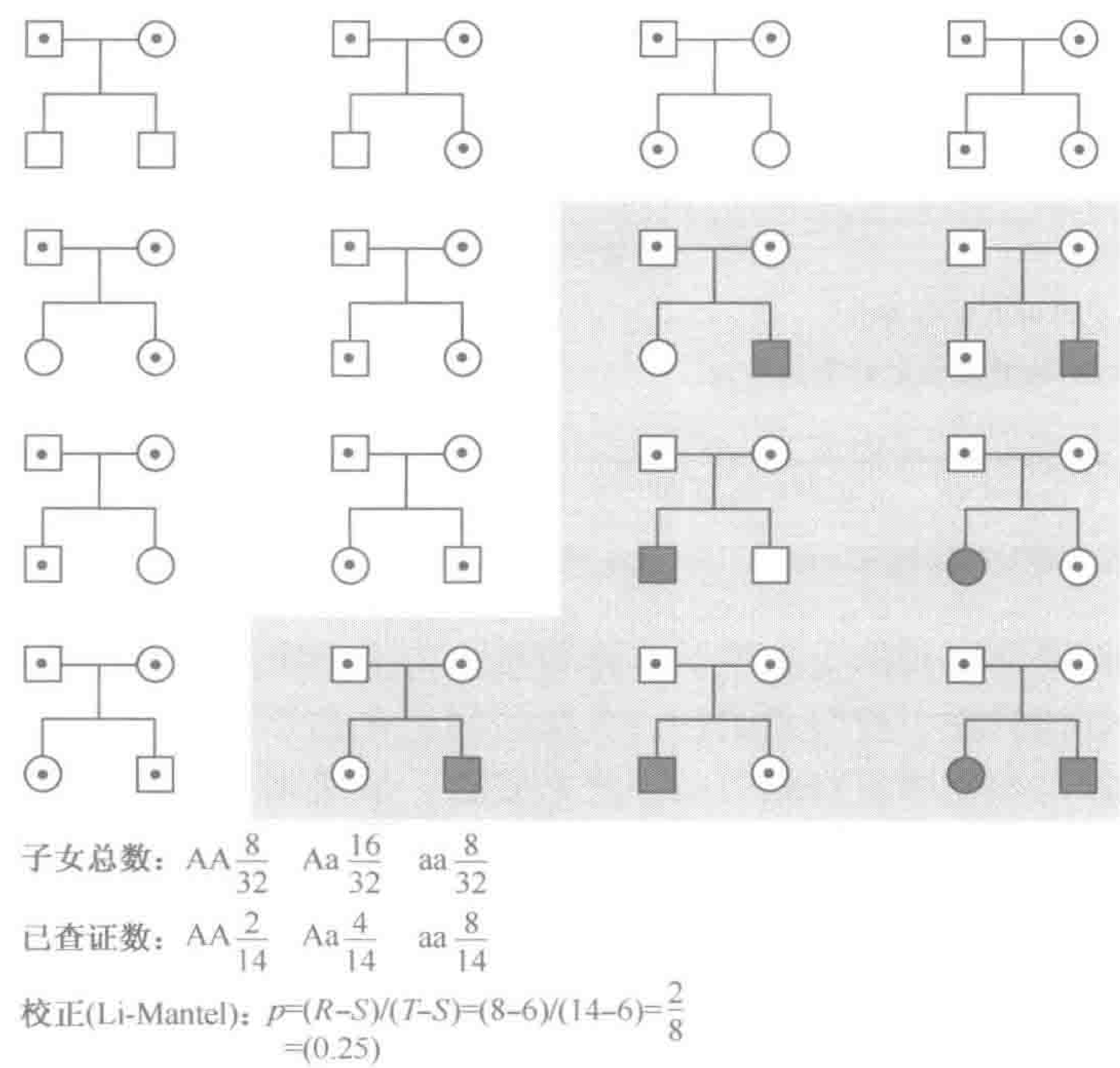


图 15.1　对某一染色体隐性疾病家系的调查偏倚（完全截短查证法）

双亲都是常染色体隐性疾病的携带者。大体上，子女 1/4 受累——但如果家系通过受累子女查证，那么只有带阴影的家系会被挑选，受累子女的比例是 8/14。通过使用 Li-Mantel 校正法可以恢复真实比例（框 15.1）。

框 15.1　校正分离率

完全截短查证法：　$p=(R-S)/(T-S)$

单一确认法：　$p=(R-N)/(T-N)$

p＝真实（无偏倚）分离率

R＝受累子女数

S＝受累独生子女数（家系中唯一受累子女）

T＝所有子女总数

N＝同胞数

上述例子以**完全截短查证法**（complete truncate ascertainment）为先决条件：我们在某一特定群体中收集了至少有一个受累子女的家系。但这并不是收集家系的唯一可行方法。我们可以通过就诊的前 100 名调查受累子女（通过长期坚持以便在同一群体中调查更多）。在这些情况下，两个受累子女的家系比一个受累子女的家系有二倍的机会被挑选，四个受累子女的家系则会有四倍的机会被挑选。**单一确认法**（single selection），被查证的机会与家系中受累子女数呈正比，出现了另一种调查偏倚，需要另一种统计学校正（框 15.1）。我们发现，要算出分离率，需要按照一种明确的查证方案收集数据，便于采用适当的校正法。

15.2.2　复杂分离分析是在大量家系数据中估计最可能参与遗传因素的一种普遍方法

对大量家族性而非孟德尔遗传疾病的病人亲属进行数据分析不是一项轻松的工作。遗传因素和环境因素可能都起作用：其中遗传因素可能是多基因的、寡基因的或孟德尔遗传方式的任意一种或任意几种的混合，而环境因素可能包括家族的或非家族变量。复杂分离分析考虑所有可能的机制、基因频率、外显率等，计算机进行最大似然性分析以找出一组参数值，该组参数值给出所观察数据的最大总体似然性。表 15.4 中显示的就是一个例子。就像用 lod 值分析（第 13 章），所提出的问题是观察数据按照一种假设比按照另一种假设有多大程度更可能成立。

表 15.4　复杂分离分析

模式	d	t	q	H	z	x	χ^2	p
混合的	1.00	7.51	9.6×10^{-6}		0.01	0.15		
散发的							334	$<1\times10^{-5}$
多基因的				1.00	1.00		78	$<1\times10^{-5}$
主要隐性基因座	0	8.22	3.8×10^{-3}				35	$<1\times10^{-5}$
主要显性基因座	1.00	7.56	1.2×10^{-5}			0.19	2.8	0.42

数据来自患有长节段先天性巨结肠病先证者查证的家系。可变参数有 t（低易感等位基因纯合子与高易感等位基因纯合子间的易患性差异，测量单位为易患性标准差）、d（任何主要致病等位基因的显性程度）、q（任何主要致病等位基因的基因频率）、H（成人中由多基因遗传引起的易患性的总体方差的比例）、z（儿童遗传率与成人遗传率之间的比率）和 x（由新突变引起的病例的比例）。编码显性易感性的单一主要基因座解释了数据以及综合模式，其中允许所有机制的混合。数据来自 Badner 等(1990)。

在表 15.4 的例子中，对特定模式（散发的、多基因的、显性的、隐性的）显示的数据与综合模式（混合模式）算出的似然性进行了比较，由于计算机可以自由优化综合模式的单基因、多基因和随机环境因素的组合。所有模式受总发病率、性别比和从收集数据估算出查证概率的制约。数据显示，单基因显性模式与混合模式无显著差异（$\chi^2=2.8$；$p=0.42$），而假定的非遗传因素模式、单纯多基因遗传模式或单纯隐性遗传模式则有差异。如果简单解释比复杂解释更适合的话，该分析提示存在患先天性巨结肠病的**主要显性易感性**（major dominant susceptibility）。现已鉴定了一些这样的基因（节 15.6.2）。

不管分离分析程序有多么聪明，也只能放大输入参数的似然性。如果遗漏了一个主

要参数，得出的结果可能会误导。McGuffin 和 Huckle 的数据很好地说明了这一点（表 15.5）。他们调查了有学医亲属的医学生。当他们把数据输入分离分析程序后，发现分析结果显著支持存在一个学医的隐性基因。尽管这结果很荒谬，但并非为好玩来做这项实验，也不是为了怀疑分离分析。作者没有让计算机考虑可能真正的原因，即共同的家庭环境。计算机下一步最佳的可能选择是数学上的效应但在生物学上是不现实的。McGuffin 和 Huckle 严肃地指出，分离分析对人类行为性状分析存在许多缺点，不审慎的分析会产生虚假的遗传效应。

表 15.5　进入医学院学医的隐性基因?

模式	d	t	q	H	χ^2	p
混合的	0.087	4.04	0.089	0.008		
散发的					163	$<1\times10^{-5}$
多基因的				0.845	14.4	<0.005
主要隐性基因座	0.00	7.62	0.88		0.11	N. S.

数据来自 McGuffin 和 Huckle(1990)对医学生及其家庭的一项调查。各符号的意义与表 15.4 相同。"患病"被定义为进入医学院学医。该分析似乎支持隐性遗传，因为它同样很好地解释了数据与非限制模式。该研究的要点在于阐述，如果共同的家庭环境因素被忽略，那么对家庭数据的分析会怎样产生以假乱真的结果(见正文)。

15.3　复杂性状的连锁分析

15.3.1　标准 lod 值分析通常不适合非孟德尔性状

标准 lod 值分析称作**参数分析**（parametric analysis），因为它需要一个精确的遗传模式、详尽的遗传方式、基因频率和每种基因型的外显率。只要有一种有效的模式，参数连锁就能提供一种极其强有力的方法，每 20Mb 进行基因组扫描以便定位疾病基因。对于孟德尔遗传性状而言，确定一个适当的模式应该不是大问题，而对于非孟德尔遗传性状则不那么容易处理。

诊断标准的重要问题

建立适用于遗传分析的诊断标准是一个主要问题。对于孟德尔遗传综合征，患者的哪些特征是综合征的一部分而哪些只是巧合常常是非常明显的。不同的特征有不同的外显率，但基本上综合征的组分是孟德尔遗传方式共分离的特征。对非孟德尔遗传疾病来说就不存在这样的实际性检验。经过不懈努力建立有效的诊断标准，尤其是对精神病研究，能够使两个独立研究小组就某一标记是否适用于某患者可以获得一致结果。但是诊断性标记可能有效而无生物学意义，尤其是对精神病和行为表型，诊断标准常常是生物学的主观性。遵循这些诊断标准有助于使不同的研究有可比性，但不能保证提出正确的遗传学问题。

对"近似-孟德尔遗传"家系的连锁分析

一旦对诊断标准达成共识，遗传分析的一条途径就是寻找近似-孟德尔（near-Men-

delian）遗传方式分离的家系。分离分析用来确定家系中遗传模式的参数，这些家系随后用于标准（参数的）连锁分析。这样的家系可能出现三种情况：

- ▶ 任何复杂疾病都很可能是异质的，所以家系收集也可能包括一些孟德尔遗传家系，其表型难以与多数的非孟德尔遗传家系区分；
- ▶ 近似-孟德尔遗传家系可能表现为大多数人碰巧都有某疾病的多种决定因素，以至于这些易感因素之一出现孟德尔遗传分离而可能打破平衡；
- ▶ “近似-孟德尔遗传”方式是“虚假”的——只是受累个体碰巧聚集在一个家系中。

在第一种情况下，鉴定其中的孟德尔遗传家系亚类有其内在价值，但无法了解非孟德尔遗传疾病的病因，乳腺癌（节 15.6.1）和阿尔茨海默病（Alzheimer disease）（节 15.6.3）就是这种情况。在第二种情况下，定位的基因座也正是常见非孟德尔遗传疾病的易感因素——先天性巨结肠病（节 15.6.2）就是例子。最后，早期研究精神分裂症的例子属第三种情况，现已普遍认为得出 lod 值为 6 是“虚假”的（Byerley，1989）。这一重大失败足以说服多数研究者转向非参数研究。

15.3.2　非参数连锁分析不需要遗传模式

无模式（model-free）的或**非参数**（nonparametric）的连锁分析方法寻找受累个体所共有的等位基因或染色体片段。Neil Risch 于 1990 年发表的三篇论文阐述了这些方法的基本理论（Risch 1990a，b，c）。**共享片段法**（shared segment method）既可用于家系（节 15.3.3），也可用于整个群体（节 15.4）。

区别片段是**传递一致性**（identical by descent，IBD）还是**状态一致性**（identical by state，IBS）很重要。等位基因 IBD 可追溯到共同祖先（通常是亲代的）等位基因的拷贝。而等位基因 IBS 看起来等同，实际上可能的确如此，但追溯不到它们的共同祖先，因此数学处理必须按照群体频率，而不是按照共同祖先传递的孟德尔遗传概率。图 15.2 说明了二者的不同。对于非常罕见的等位基因，不太可能有两个相互独立的起源，所以一般 IBS 暗示 IBD，但不适用于常见的等位基因。对于确定 IBD，多等位基因微卫星比双等位基因标记更有效，而多基因座的多等位基因单体型则更好，因为任何一种单体型都很可能是罕见的。只要运用恰当，IBS 数据或 IBD 数据都可以用于共享片段分析。IBD 更为有效，但需要更多亲属的样本。对于一个有数个受累个体的复杂系谱，可以用标记信息来计算一对受累亲属共享单体型 IBD 的概率（Arnos *et al.*,1990）。

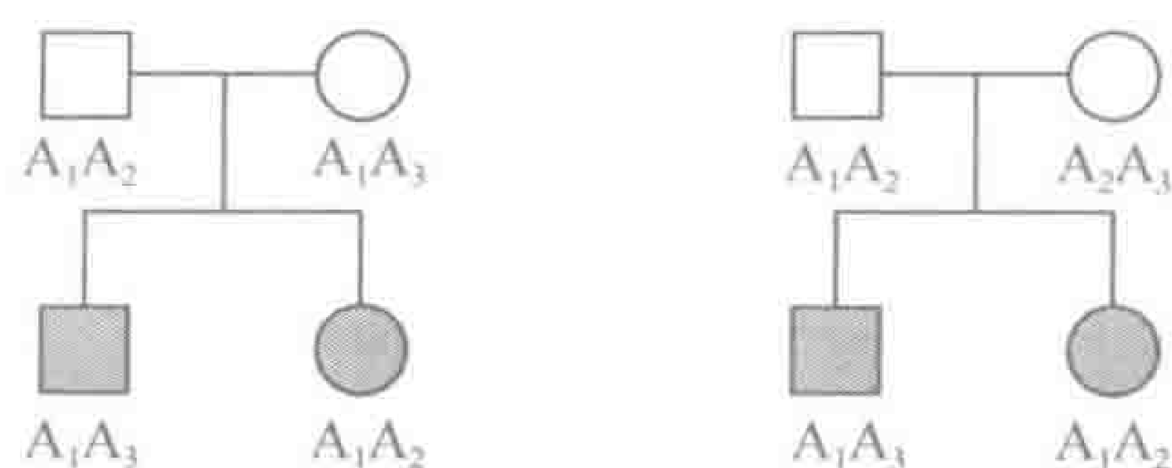


图 15.2　状态一致性（IBS）和传递一致性（IBD）

两个同胞对都共享等位基因 A_1。第一个同胞对有 A_1 两个独立拷贝（IBS 而不是 IBD）；第二个同胞对共享同一个亲代 A_1 等位基因拷贝（IBD）。只有在亲代基因型已知的情况下才能明显区分。

15.3.3　家系共享片段分析：受累同胞对和受累家系成员分析

随机选择一段染色体，同胞对预期共享 0、1 或 2 个亲代单体型的频率分别为 1/4、1/2 和 1/4。而如果一对同胞都患有一种遗传性疾病，那么他们很可能共享带有疾病基因座的染色体片段。如果每个患者的基因座都携带某个突变等位基因，那么，若该病为

显性的话，这对同胞至少共享一个亲代单体型，若该病为隐性的话，他们至少共享两个亲代单体型（图 15.3）。这允许进行简单形式的连锁分析。**受累同胞对**（affected sib pair，ASP）以标记分型，寻找染色体区域共享 2、1 或 0 个单体型传递一致性，其随机比率高于 1：2：1。如果只是检测同胞对状态一致性，则必须按无效假设以基因频率函数计算预期共享。进行 ASP 分析不需要对疾病的遗传学作任何假设，并且受累同胞对收集比扩展家系通常更容易。**多点分析**（multipoint analysis）比单点分析更优越，因为能更有效地获取染色体区域共享 IBD 的信息。Kruglyak 和 Lander（1995）的 MAPMAKER/SIBS 程序被广泛地用于分析多点 ASP 数据，得出**非参数** lod（nonparametric lod，NPL）值。

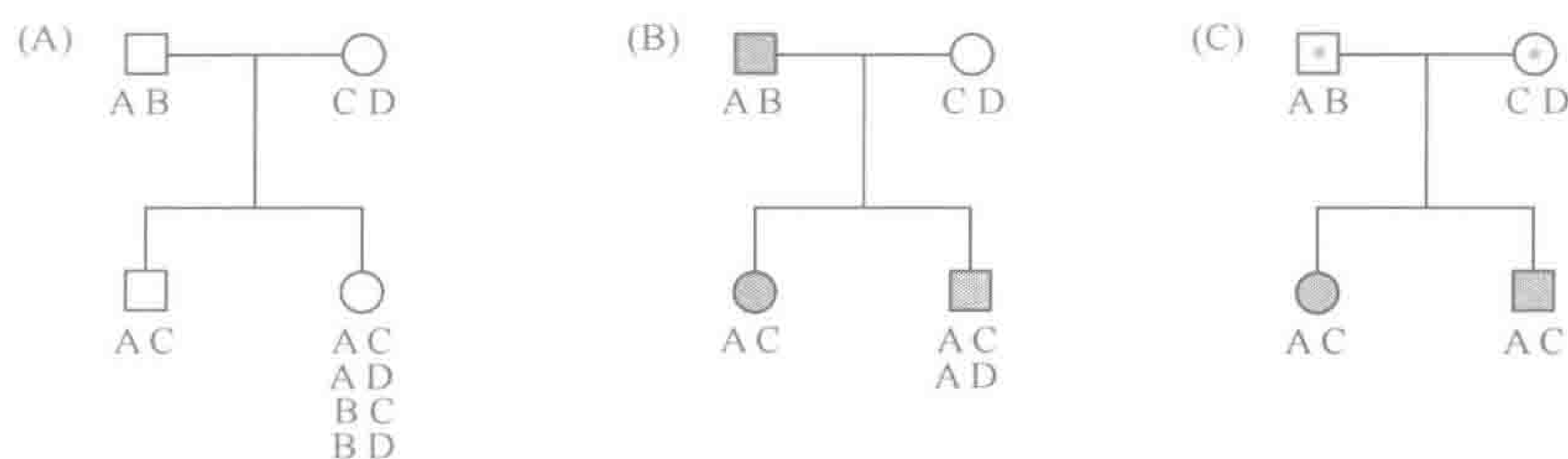


图 15.3　受累同胞对分析

（A）通过随机分离，同胞对共享 0、1 或 2 个亲代单体型的概率分别为 1/4、1/2 和 1/4。（B）同胞对都患有一种显性疾病，则共享相关亲代染色体片段的一个或两个拷贝。（C）都患有一种隐性疾病的同胞对必然共享两个相关染色体片段的亲代单体型。受累同胞对的上述随机单体型共享可以鉴定包含易感基因的染色体片段。

ASP 分析的一个缺陷是为了定位克隆所鉴定的候选区域进行通常不能太大。不管是巧合还是因为共享易感性用少数重组体来鉴别同胞，以至于同胞共享亲代大片段染色体。重要是复杂疾病分析过程不像孟德尔遗传基因定位最终结果那样，检测标记越来越近直到不再可能重组。如果一个易感基因既不必需也不足以致病，那么不是所有的受累同胞对都会共享这一相关染色体片段。此外，同胞对会巧合地共享许多片段。然而，ASP 基因定位由于其简单性和稳定性，一直是寻找常见非孟德尔遗传疾病易感性基因的主要手段之一（节 15.6）。Sham 和 Zhao 详述了 ASP 分析的数学计算（进一步阅读）。

诸如 GENEHUNTER 之类的程序（节 13.6.2）将共享片段分析扩展到其他亲属关系。这些程序计算受累亲属共享等位基因传递一致性的程度，并且将所有受累家系成员的计算结果与单纯孟德尔遗传分离的无效假设进行比较（遗传标记遵循孟德尔分离率，除非被连锁或关联违背）。这种比较可以用于由计算机计算出非参数 lod 值。

15.3.4　显著性阈值是复杂疾病分析应考虑的重点

通过显著 lod 值定位的绝大多数孟德尔遗传基因已被成功克隆，然而复杂疾病分析历程却展示了一系列不可重复的结果。各项独立研究极少一致地确定某一疾病的候选区域。Risch 和 Botstein（1996）概述的躁狂－抑郁精神病研究的典型经历，以及 Altmüller 等（2001）综合分析的 31 种复杂疾病 101 项连锁研究都证明了这一点。不管

每一项研究中导致这些问题的真正原因是什么，有一个显而易见的共性就是很难确定什么时候称此结果是显著的。

确定合适的显著性阈值这一问题既有几分专业性又有几分哲学性。我们已经注意到**逐点**（pointwise）（或**正态**，nominal）**显著性**（significance）和**全基因组显著性**（genome-wide significance）之间的不同（节 13.3.4）。

- 设无效假设为无连锁，连锁统计的**逐点** p **值**是基因组某一特定位置超过观察值的概率；
- 设无效假设为无连锁，**基因组** p **值**是基因组任何位置超过观察值的概率。

对于全基因组研究，恰当的显著性阈值就是基因组任何位置发现假阳性概率为 0.05 时的数值。理论上提示（Lander and Kruglyak，1995）受累同胞 IBD 检测其基因组 lod 值的阈值为 3.6，IBS 检测其阈值为 4.0。复杂疾病研究通常通过模拟来估计它们的显著性阈值，显然，计算机是通过随机标记基因型而产生家系收集的 1000 个模拟，但要依据正确的等位基因频率、重组率等。每个模拟的数据组都进行了全基因组搜索并标明了最大 lod 值。基因组的显著性阈值是超过小于模拟的 5%的数值。

针对疾病易感基因定位重复实验频频失败的现象，Lander 和 Kruglyak（1995）提出了表 15.6 中的一系列阈值。注意逐点 p 值 1×10^{-5} 并不等同于 lod 值 5.0——两种量度并不相同：

- lod 值为 5.0 的含义是既定连锁假设比无效假设的数据多 10^5 倍的可能性；
- p 值为 10^{-5} 意味着既定无效假设所确定的 lod 值有超过 10^5 之一的可能性。

关于 Lander 和 Kruglyak 标准的一些讨论，参见 Nature Genetics 1996 年 4 月刊的相应部分。

表 15.6　表示连锁的建议标准（Lander and Kruglyak 1995）
数字 p 值和 lod 值来自 Altmüller 等（2001）

连锁类型	全基因组扫描巧合发生的预期次数	p 值的大约范围	lod 值的大约范围
暗示连锁	1	$7\times10^{-4}\sim3\times10^{-5}$	2.2～3.5
显著连锁	0.05	$2\times10^{-5}\sim4\times10^{-7}$	3.6～5.3
强烈显著连锁	0.001	$\leqslant3\times10^{-7}$	$\geqslant$5.4
确定连锁	0.01 在前一次独立研究有显著连锁的某候选区域		

15.4　关联研究与连锁不平衡

关联（association）并不是一种特别的遗传学现象；它只是等位基因或表型共发生的一种统计学表述。如果某些人同时有疾病 D 和等位基因 A 的频率常高于（或/也可以常低于）群体中预计 D 和 A 各自的频率，那么等位基因 A 与疾病 D 关联。例如，在英国发现普通群体中有 HLA-DR4 的占 36%，而类风湿性关节炎的群体中有 HLA-DR4 的占 78%。

15.4.1 为什么会发生关联

一个群体关联可能有很多原因，并非都是遗传性的：

- ▶ **直接原因**：有等位基因 A 使人易感疾病 D。拥有 A 可能既不必需也不足以使某人患疾病 D，但增加了这种可能性；
- ▶ **自然选择**：那些患有疾病 D 的人如果也有等位基因 A，则很可能生存下来并生育子女；
- ▶ **群体分层**：群体包含几种遗传上不同的亚群，疾病 D 和等位基因 A 同时常见于一个亚群中。Lander 和 Schock（1994）举出旧金山海湾地区 HLA-A1 与能用筷子吃饭关联的例子。HLA-A1 在中国人中比在白人中更常见；
- ▶ **Ⅰ类错误**：关联研究通常检验大量标记与某一疾病的关联。即使没有任何真正的作用，5%的结果会有显著性在 $p=0.05$ 水平，1%为显著性在 $p=0.01$ 水平。原始 p 值需要根据提出的问题进行校正（节 15.4.4）。过去，研究者常没有进行充分的校正，所报道的关联不能在后继的研究重复；
- ▶ **连锁不平衡**（LD）：复杂疾病关联研究的目的在于发现标记和疾病间 LD 引起的关联。下面将讨论 LD 现象，框 15.2 描述了 LD 如何检测。

框 15.2 连锁不平衡的检测

如果两个基因座有等位基因 A、a 和 B、b，其频率分别为 p_A、p_a、p_B 和 p_b，会有四种可能的单体型 AB、Ab、aB 和 ab。设这四种单体型的频率为 p_{AB}、p_{Ab}、p_{aB} 和 p_{ab}。如果没有 LD，那么 $p_{AB}=p_A p_B$，以此类推。背离这种随机关联的程度可以用 $D=p_{AB}p_{ab}-p_{Ab}p_{aB}$ 来衡量。

作为 LD 的一个量度，D 有个缺点，即其最大绝对值取决于两个基因座的基因频率，以及不平衡的程度。首选的量度有：

- ▶ $D'=(p_{AB}-p_A p_B)/D_{max}$，$D_{max}$ 是既定等位基因频率条件下 $|p_{AB}-p_A p_B|$ 可能的最大值。
- ▶ $\Delta^2=(p_{AB}-p_A p_B)^2/(p_A p_a p_B p_b)$

D' 应用最广泛。它在 0（无 LD）和 ±1（完全关联）之间变化，而且比 D 较少依赖于等位基因频率。凭经验估计，$D'>0.33$ 常看作 LD 的阈值水平，高于此值时在一般大小数据组会看到明显的关联。其他检测方法的增多提示了没有一个检测是理想的（Devlin and Risch，1995）。尤其是，这些检测都是为配对基因座而开发，而绝大多数全基因组扫描采用多点分析。这样的数据应该用于检查保守单体型，而不只单用于配对 LD 分析。

15.4.2 理论上关联截然不同于连锁，而家系和群体相容了，连锁和关联也相容了

在理论上，连锁和关联是完全不同的现象。连锁是基因座之间的关系，而关联是等位基因间或表型间的关系。连锁是一种特殊的遗传学关系，而关联，如上面所说，单纯是一个对各种因素的统计学观察。

连锁本身在一般群体中不产生任何关联。例如，*STR45* 标记基因座与抗肌萎缩蛋白（dystrophin）基因座连锁，然而，*STR45* 等位基因在无亲属关系的 Duchenne 肌营

养不良患者中的分布与在一般群体中的分布完全一样。当一个家系中 dystrophin 基因突变出现分离，我们可以预料到受累个体共享相同的 *STR45* 等位基因，因为这两个基因座是紧密连锁的。这样，连锁在家系中可以产生关联，但在无亲属关系的则不能。然而，如果患疾病 D 的两个个体假设无亲属关系，实际上是从一个遥远的共同祖先遗传下来的这种疾病，那么他们也可能倾向于共享特定祖先等位基因，并且该基因座与疾病 D 紧密连锁。节 13.5.2 显示了这一现象的例子。

共同祖先很重要，因为我们所有人都有。只要我们追溯到很远，整个人类都有关系。如果到目前为止，一个群体就是一个扩展的家系，那么，由于 LD，祖先的疾病易感基因和与其紧密连锁的标记间应该存在群体水平上的关联。一项粗略的计算提示，在英国两个“无亲属关系”的个体至少在 22 代以前拥有共同的祖先。如果是完全远亲后代，此期间他们每人将会有 $2^{22}\approx$400 万个祖先。22 代大约是 500 年，在 1500 年英国人口数大约就是 400 万（图 15.4）。

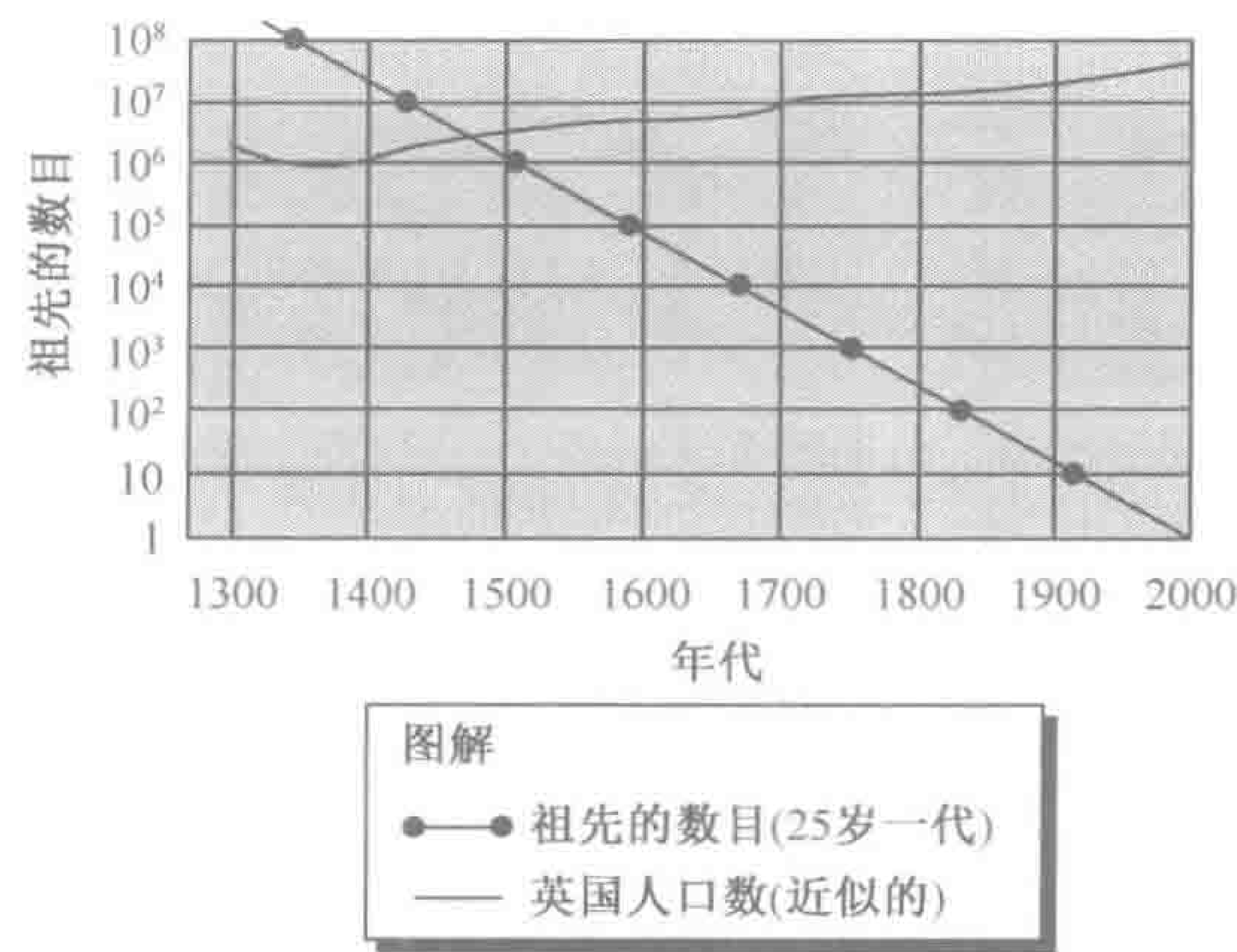


图 15.4　渐渐融合到基因库

一个完全远亲杂交的个体在 n 代之前有 2^n 个的祖先，如果英国群体都完全是远亲杂交后代，两个“无亲属关系”的现代人会有完全一样的 1500 年的祖先。当然现实上，英国群体并不完全是远亲杂交后代，那两个人会有强烈重叠但不完全相同的 1500 年的祖先群。

假如两个“无亲属关系”的个体各自都从他们的共同祖先遗传了一种疾病的易感等位基因，许多代和许多次减数分裂使他们与共同祖先有别，反复重组将把共享的染色体片段减小至非常小的区域，只有与疾病易感基因座紧密连锁的等位基因才仍可共享。当一个易感基因座的重组率是 θ 时，每一代会有 θ 比率的祖先染色体失去关联，有（$1-\theta$）的比率保留关联。n 次减数分裂后，$(1-\theta)^n$ 比率的染色体会保留关联。基因座 1cM 和 2cM 的 LD 半衰期分别为 69 次减数分裂和 34 次减数分裂，因为 $(0.99)^{69}\approx(0.98)^{34}\approx0.5$。我们上面算出，“无亲属关系”的两个英国人的祖先在 22 代以前完全融合为一个。这种计算略简化些，因为假定了整个英国群体在过去的 500 年间是一个自由相互交配的社会。然而，它得出了一个最初的原始估计，即在英国群体中，等位基因关联反映了共享祖先片段，并且开始注意了相距 1cM 以内的基因座。更成熟的计算是使用重组事件的泊松分布，并结合了有关群体结构和历史的假设（Kruglyak，1999）。

一个关键的决定因素是**融合时间**（coalescence）——追溯到最近共同祖先的世代数（框12.6）。然而，广泛而随机的变异以及未知的群体历史细节使人们对最精确的计算也不相信。我们需要的是数据，最近已经可以获得越来越多的真实数据了。

15.4.3 许多研究显示了连锁不平衡岛被重组热点分开

囊性纤维化基因通过 LD 梯度上升达到最大值而被鉴定（节 13.5.2）。然而，对其他疾病的研究很快发现，那种 LD 顺畅梯度只是例外而非正常的。对 Huntington 病（图 15.5）我们可以发现疾病与较远的标记有强关联，而与较近的标记有弱关联。甚至更奇怪的是，由于 *HD* 基因座紧密连锁的 *D4S95* 标记可以用三种酶 *Taq* Ⅰ、*Mbo* Ⅰ和 *Acc* Ⅰ检测 RFLP，一些独立研究结果证实，该病与 *Acc* Ⅰ和 *Mbo* Ⅰ等位基因强烈关联，但与 *Taq* Ⅰ等位基因无关联。这样令人迷惑不解的模式反映了一个复杂历史，即小建立者群体发生了随机重组事件，或者一些标记多态性的起源可能比一些疾病的突变更近一些。

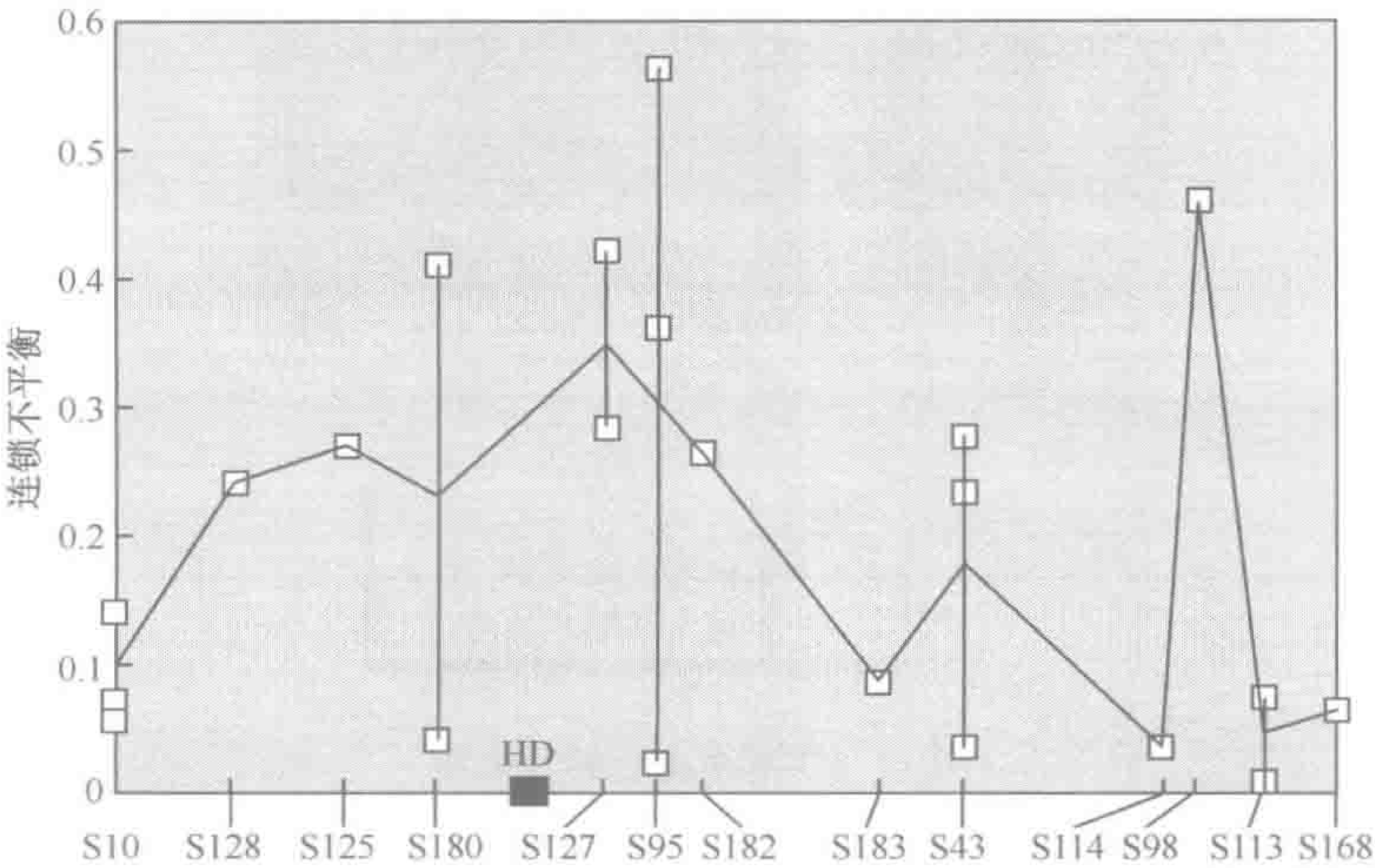


图 15.5　亨廷顿病（Huntington disease）基因座附近的连锁不平衡

S10、S125 等是 DNA 标记 *D4S10*、*D4S125* 等的缩写，显示了它们相对于 *HD* 基因座的图距。总距离代表 2500kb。对于一些基因座，存在几个不同的 RFLP，有时这些 RFLP 显示非常不同的等位基因关联，例如标记 S95（见正文）。引自 Krawczak and Schmidtke（1998）*DNA Fingerprinting*，2nd ed. BIOS Scientific Publishers，Oxford。

最近，大量的系统研究报道了由重要染色体片段标记-标记的 LD（Gabriel *et al*.，2002b 和其中的参考文献）。一个共同发现是，LD 并不随距离缩短而平稳减小。相反，染色体含有一系列相对长的 LD 岛，这些 LD 岛相互间明显地隔开（图 15.6）。在岛内部，有用的 LD 可延伸 50kb（指欧洲人；非洲人的小一些），但不同的岛，即使是间隔很近的标记相互间也没有 LD。对一些区域的详细检测证实了岛的边界确实是重组的热点。这大概为我们展示了，祖先染色体片段的嵌合体构成我们共同的遗产。如果能够确定一个群体 LD 岛在全基因组的结构，那么就可能确定一组标记（“hap-SNP”）来建立每一个岛的单体型，这些单体型又可以用于检测疾病的关联。有人认为，在任何一个群

体中，绝大多数岛只会有 4～6 种不同的常见单体型（Gabriel *et al*.,2002b）。目前正进行大规模的标记-标记定位来确定这些结构（http://www.genome.gov/10005336）。

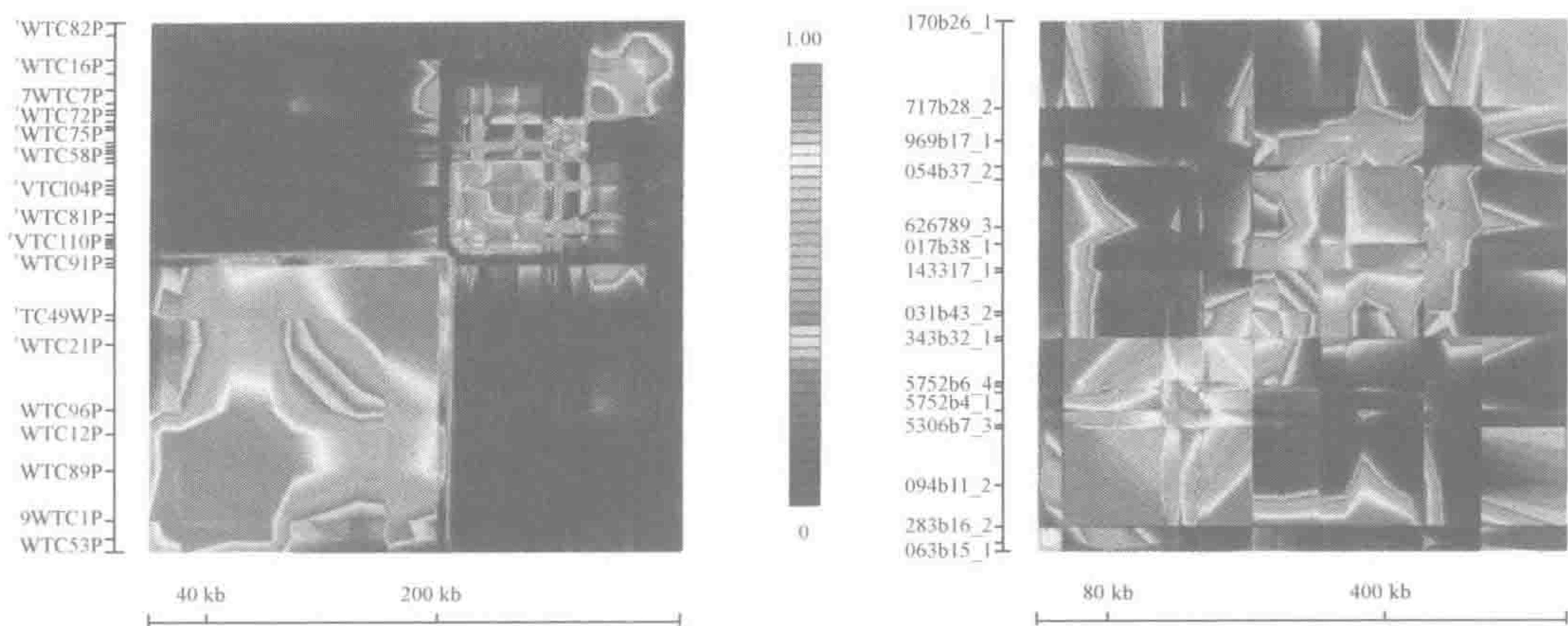


图 15.6　两个染色体区连锁不平衡的方式

每一个正方形中，同组标记按染色体顺序在 X 轴和 Y 轴上表示。按照中间的标尺，每一对标记在 Cartesian 坐标上的颜色显示了 LD 的强度。GOLD 程序通过内插法填补了点和点之间的空隙。如果 LD 只是距离的作用，那么每一个方形在对角线一致是红色，各点远离对角线逐渐变为一致的蓝色。需注意的是独立的 LD 岛的基本形式，但有很多复杂的细节。左图为 2 号染色体部分；右图为 13 号染色体部分。本图由 Dr. William Cookson，Oxford 提供。

15.4.4　关联研究的设计

寻找群体关联对于鉴定疾病易感基因是一个诱人的选择。关联研究比连锁分析更容易进行，因为它不需要多发病例的家系或特殊的家系结构。在有些情况下，关联检测弱易感等位基因比连锁更有效力（见下文）。然而，最重要的是要认真考虑实验设计。

关联检测方法的选择

任何关联研究的对照组选择都是非常关键的。要尽可能使对照与病例相匹配，绝对没有疏漏也是不可能的。因此，当发现一种关联时，总会担心关联是由不充分匹配的对照引起的，而不是由与易感基因座的连锁不平衡引起的。这种不确定性的叠加以及过多不可重复的结果（尤其是有关 HLA-疾病的关联），使得病例-对照研究在 20 世纪 80 年代期间受到人类遗传学家的冷落。

最近，研发了一些能大大防止这类问题发生的方法。这些方法可以称为**内对照的关联研究**（association study with internal control）。最流行的方法是**传递不平衡检验**（transmission disequilibrium test，TDT；Schaid，1998）。TDT 检测有一个或多个患病子女的夫妇，它与双亲之一是否患病无关。为了检测标记等位基因 *M1* 是否与疾病关联，我们选择 M1 杂合子的亲代。该检验是比较亲代传递 *M1* 等位基因给患病后代的病例数与亲代传递另一个等位基因的病例数（框 15.3）。其结果不受群体分层的影响。有人研发了一种扩展 TDT（ETDT；Sham and Curtis，1995），用来处理如微卫星的多等位基因标记的数据。TDT 可以用于只有一个亲代资料的情况，但此时会因偏倚而影响结果（Schaid，1998）。当没有亲代资料时（晚发疾病的一个常见问题），另一种变型 TDT——

同胞-TDT 可用来观察受累与未受累同胞间标记等位基因频率的不同（Spielman and Ewens，1998）。

框 15.3　传递不平衡检验（TDT）检测标记等位基因 *M1* 是否与疾病相关

1）受累先证者被查证

2）先证者及其双亲被按标记分类

3）挑选标记等位基因 *M1* 是杂合子的双亲，他们可能受累也可能不受累

设 a 为杂合子亲代将 *M1* 传递给受累后代的次数，b 为另一个等位基因被传递的次数。

TDT 的检测统计量是$(a-b)^2/(a+b)$，假如数目适当大，则服从 χ^2 分布，自由度为 1。

曾有一些关于 TDT 是连锁检测还是关联检测的争议。因为它提出的问题是关于等位基因而不是基因座，所以基本上认为是一种关联检测。关联的等位基因本身可以是易感基因，或者是与附近基因座的易感等位基因处于连锁不平衡。如果没有连锁不平衡，TDT 就不能检测连锁——关键是记住什么时候考虑应用 TDT 进行全基因组扫描。

传统的病例-对照研究作为另一个不同于 TDT 的选择现在又受到青睐。病例-对照研究比 TDT 需要更少的样本，且对缺乏亲代资料的晚发疾病更容易些。有人认为夸大了由于群体分层造成假关联的危险性，而且可以通过比较无连锁基因座等位基因频率的病例和对照研究来检查数据是否存在可能的分层效应（Pritchard and Rosenberg，1999）。按照 Risch 和 Teng（1998）的说法，以受累同胞对作为病例，并设两个无亲属关系的人为对照。

不管用什么关联检测，一定会提出多重检测的问题。当 N 值很大时，完全 bonferroni 校正是过于保守了（将 p 阈值被 N 除，N 为提出问题的总数）。理想的显著性阈值为$p'=1-(1-p)^N$（Emahazion *et al*.,2001）。Cardon 和 Bell（2001）对关联研究设计的一些普遍问题进行了非数学推理性的论述——见“进一步阅读”。

标记的选择

单核苷酸多态（SNP，框 13.1）是关联研究选择的标记，有两个理由：

▶ 不像微卫星那样，SNP 的数量（平均每 kb≥1）足以确定 LD 岛，而且可以用各种高通量的方法来统计（节 18.4.2）。

▶ SNP 比微卫星突变少。如果这一点确切的话，正如常提示的那样，常见病的易感性主要是由共同祖先的 DNA 变异决定的，利用长期保持稳定的标记来鉴定祖先的单体型。

对于已知连锁不平衡模式（如图 15.6 中阐述的）的染色体区域和群体，可以选择标记来确定在每一个 LD 岛中的单体型。没有这些信息，一切都是猜测。人类历史上一个疾病的等位基因越古老，就需要越高密度的 SNP 来检测这个等位基因（Kruglyak，1999；Wright *et al*.,1999）。有些研究者喜欢利用位于基因内的 SNP，尤其是编码序列的 SNP（cSNP），因为他们认为这些变异更可能是真正的易感性决定因素。实际上，目前没人知道标记选择的最适策略。不同的群体中的不同疾病有不同的历史，而每一项研究都需要一个拟定策略。

研究群体的选择

一个争论问题是关联研究在隔离群体中是否会更有结果。预期由少数建立者衍生的群体会显示出有限的单体型多样性和较高的连锁不平衡。相信易感基因座的致病等位基因在隔离群体中更容易鉴定的理由在于冰岛的解码项目（Gulcher *et al*.,2001）以及其他类似方案。对于有群体特征的罕见孟德尔遗传疾病，这样的群体会充分显示出基因座附近强烈而大范围的连锁不平衡［例如，芬兰的芬兰病（Finnish disease)]，而这种不平衡只存在于携带疾病等位基因的染色体，该疾病等位基因大约都起源于一个共同祖先。对于更常见的变异，实验数据并没有明显地显示增加不平衡（Varilo *et al*.,2000；Pritchard and Przeworski，2001)。获得有完善医疗记录的大量受试者可能比群体结构更重要——当然这种想法源自于英国的 BioBank 计划，该计划旨在收集500 000名 45～69 岁英国人的医疗和生活数据以及 DNA，并追踪他们将来的健康状况。

撒哈拉以南地区的非洲群体比欧洲群体有更高的遗传多样性，这与人类起源于非洲的假说相符，且有限的数据提示非洲人的连锁不平衡是短范围的（Reich *et al*.,2001)。另一方面，源自于近代混合人种的群体显示非常强烈的连锁不平衡（例如，Wilson 和 Goldstein（2001）研究的 Lemba 群体是 Bantu-Semitic 的杂交群体)。理论上，如果两种群体在某种常见病的发病率上有很大不同，那么其混合群体就可以有效地用于定位该病的决定因子（就像在小鼠杂交实验里一样)，但这种想法尚未在实际中检验。总而言之，目前还不清楚某群体是否在复杂疾病关联研究方面有特殊优势。有关这些问题的深入讨论，见 Wright 等（1999）和 Peltonen 等（2000)。

基因型还是单体型?

在研究个体而不是家系时，原始数据是基因型构成，而关联分析需要的是单体型。单体型可以通过计算机分析最大期望值而由基因型推断出来（Long *et al*.,1995)，但理论上它不可能达到 100%的可信度；唯一全面而可靠的方法是分析含有单倍体染色体的杂交体细胞（Douglas *et al*.,2001)。有人认为这是目前大规模关联研究设计的基本缺点。乐观人士相信关联研究会奏效，是因为大多数 LD 岛（节 15.4.3）只含有少数常见的单体型，他们坚信从基因型中可以鉴定出这些单体型。有意思的是，看谁是对的。

连锁还是关联：玩数字游戏

Risch 和 Merikangas（1996）发表的一篇重要论文指出，关联检测弱易感等位基因比连锁更有效。他们鉴定了与疾病易感基因座紧密连锁的标记，比较了连锁（受累同胞对，ASP）和关联（TDT）检测的效力。在既定效能和显著性水平的情况下，他们计算了区分遗传学效应与无效假设所需的 ASP 数或 TDT 核心家系（受累子女和双亲）数。框 15.4 阐述了他们的方法（更多详细信息，请查阅原文)，表 15.7 显示了运用他们的公式得到的典型结果。结论很清楚，ASP 分析需要大样本数通常难以达到，才能检测出相对风险低于 3 的易感基因座，而 TDT 用可行的样本量就可检测相对风险低于 2 的等位基因。两种方法都很难找到使疾病相对风险低于 1.5 的易感等位基因。然而，值得注意的是，他们的结果掺入了多种假设——尤其是假设了疾病基因座只有一个祖先

易感等位基因。任何等位基因异质性都会迅速破坏关联检测的进行。有一些数据正说明这一点（节 15.6.6）。

框 15.4 检测疾病易感基因座所需的样本大小——通过应用受累同胞对（ASP）或传递不平衡检验（TDT）进行全基因组扫描

Risch 和 Merikangas（1996）计算了在效能（$1-\beta$）和显著性水平 α 情况下区分一种遗传学效应和无效假设所需的样本数。本框概括了他们的公式及方程式，但要了解其衍生和详细信息，应查阅原文。

一项标准统计量告诉我们，所需的样本大小 M 由 $(Z_\alpha-\sigma Z_{1-\beta})^2/\mu^2$ 得出，这里 Z 指的是标准正态偏差。均数 μ 和方差 σ^2 作为易感等位基因频率（p）和由一个易感等位基因所致的相对风险度 γ 的函数计算得出。该模式假定携带两个易感等位基因的个体的相对风险度是 γ^2；假定所用的标记总是能提供信息；并且假定易感基因座没有重组。

对于 ASP，易感基因座的预期共享等位基因由 $Y=(1+w)/(2+w)$ 得出，其中 $w=[pq(\gamma-1)^2]/(p\gamma+q)$。$\mu=2Y-1$，$\sigma^2=4Y(1-Y)$。基因组范围显著性阈值（基因组任何位置出现假阳性的概率＝0.05；检测 IBD 共享）需要 lod 值为 3.6，相应的 $\alpha=3\times10^5$，$Z_\alpha=4.014$。对于 80%的效能来检验效应，$1-\beta=0.2$，$Z_{1-\beta}=-0.84$。

对于 TDT，一个亲代为等位基因的杂合子的概率是 $h=pq(\gamma+1)/(p\gamma+q)$。$P$（trA），即这样一个杂合子亲代把高风险度等位基因传递给受累子女的概率为 $\gamma/(1+\gamma)$。$\mu=\sqrt{h}(\gamma-1)/(\gamma+1)$，$\sigma^2=1-[h(\gamma-1)^2/(\gamma+1)^2]$。如上面讨论过的，对于一项1 000 000次检测的基本基因组扫描，$\alpha=5\times10^{-8}$，$Z_\alpha=5.33$，和前面一样，$Z_{1-\beta}=-0.84$。

表 15.7 中，Z_α，$Z_{1-\beta}$，μ 和 σ^2 通过带入公式 $M=(Z_\alpha-\sigma Z_{1-\beta})^2/\mu^2$ 计算样本大小。对于 TDT，计算结果要减半，因为每一个双亲-子女核心家系可允许两次检测，即每个双亲一次。

表 15.7 在全基因组扫描中，80%效能检测显著的连锁或关联所需的样本大小

γ	p	ASP 分析		TDT 分析	
		Y	N-ASP	P(trA)	N-TDT
5	0.01	0.534	2530	0.830	747
	0.1	0.634	161	0.830	108
	0.5	0.591	355	0.830	83
3	0.01	0.509	33797	0.750	1960
	0.1	0.556	953	0.750	251
	0.5	0.556	953	0.750	150
2	0.1	0.518	9167	0.667	696
	0.5	0.526	4254	0.667	340
1.5	0.1	0.505	115537	0.600	2219
	0.5	0.510	30660	0.600	950
1.2	0.1	0.501	3951997	0.545	11868
	0.5	0.502	696099	0.545	4606

γ 是基因型为 Aa 与 aa 的个体相比的相对风险度；p 是易感等位基因 A 的频率。对于受累同胞对（ASP）分析，Y 是预期共享等位基因，N-ASP 是基于 IBD 检测（$\alpha=3\times10^{-5}$）为达显著性所需的同胞对数。对于传递不平衡试验（TDT），P(trA)是一个 Aa 亲代把 A 传递给受累子女的概率，N-TDT 为达显著性所需要的双亲-子女核心家系数。依据为 Risch 和 Merikangas(1996)。

15.4.5　连锁和关联：互补的方法

连锁和关联在很多方面提供互补的数据。连锁可以在染色体大范围操作并且能通过几百次检测扫描整个基因组。一项代表性的研究即 250 个受累同胞的 300 个标记需要产生 $1.5\times10^5\sim3\times10^5$ 个基因型（取决于亲代是否能基因分型）。对于一个高度自动化且资金充裕的实验室，这项工作几个星期就能完成。另一方面，我们已知连锁不平衡是发生在小范围内的现象，典型的 LD 岛的大小为 20～50kb。假设 LD 岛的大小平均为 25kb，对 300 个核心家系（受累子女及其双亲）进行一次全基因组 TDT 扫描，即使每一个岛只检测一个 SNP，也需要 10^8 种基因型。现代技术也许很快会提供这样的高通量，但花费依然会令人却步。因此关联研究必须聚焦在预定的候选区域。可以参考动物模型或已知基因提示的区域，而另一选择是可以连锁分析确定的。如前所述，ASP 研究确定的候选区域对定位克隆易感基因通常太大，因此一项自然的研究设计是通过连锁，可能是受累同胞对间的连锁，开始全基因组筛查，然后，一旦得到一个初步定位，就通过连锁不平衡定位来缩小候选区域。

15.5　鉴定易感等位基因

对于孟德尔遗传疾病，发现其确切的基因可能会很困难，但一旦有人获得成功，这个基因通常是很明显的。连锁分析紧密定位的基因，在患者有明确的突变，而对照没有。对于复杂疾病总体来说，难以区别真正的易感性因素还是无关的 DNA 多态性，这有三点原因：

- 没有单一基因突变是必需或足以导致疾病的，因此，即使是真正的易感等位基因也会在一些对照中发现，而在一些患者中缺乏。此外，不同的群体中易感性的主要决定因素也可能会不同；
- LD 的补丁样特性，即大范围相关与小范围不相关共存。这意味着，尽管理论上关联研究有很高的分辨率，但实际上也不能在正确的位置上寻找易感性决定因素。此外，也没有一种遗传学方法能从相互处于强烈连锁不平衡的一系列等位基因中鉴定出真正的决定因素；
- 引起常见病易感性的遗传变异可能并不是显而易见的突变。尽管偶有例外，孟德尔遗传疾病通常是突变导致基因完全失活或至少严重影响其表达所致（第 16 章），而又通常难以确定一个候选 DNA 序列的变异是否如此。常见病的易感性很可能是一些基因表达微小改变的组合效应，并非其中一个单独致病，而这些可能在健康群体中相当常见。易感性用数量性状基因座（QTL，节 15.6.8）作为模型比二元（存在/缺失）因素更好。易感性变异在非编码 DNA 中可能是多态性，微弱地影响启动子活性、剪接或 mRNA 稳定性，例如 2 型糖尿病易感性有关的 UCSNP-43G 等位基因（节 15.6.5）位于候选基因内含子内，在未受累对照中的频率为 0.75。

关于这一问题的深入讨论（以 2 型糖尿病为背景）参见 Altshuler，Daly 和 Kruglyak（2000）的综述。

15.6 复杂疾病遗传剖析取得不同程度成功的8个例子

复杂疾病遗传学分析没有使之一致的情况。我们不打算概括整个领域的现状——每一种疾病都不同。读者对某种疾病感兴趣，应该用 PubMed 锁定一篇该疾病的最新综述。这里选择 8 种疾病来阐明复杂疾病研究这一主题。我们并不注重展示成功的事例。有些病例几乎没有进展。因为对每一个病例倾注了大量心血，没有成功也几乎如同成功一样有意义。

15.6.1 乳腺癌：鉴定一种孟德尔遗传亚类促进了重要的医学进步，但并不能解释常见散发乳腺癌的病因

尽管常见的癌症通常是散发的，但已知“癌家族”的存在有很多年了。当几个亲属都患有同一种罕见的癌，如前庭神经鞘瘤（NF2，MIM 101000）时，很容易被怀疑是一个孟德尔遗传综合征，这种家系的研究已鉴定了一个肿瘤抑制基因，如节 17.4 所述。乳腺癌很常见——英国妇女一生的患病风险度是 1/12——因此当几个亲属都患有乳腺癌时，就不清楚是运气不好还是一个真正的癌家族。然而，明显的乳腺癌家族史通常与异常早发、乳腺癌合并卵巢癌、双侧肿瘤多见以及偶有男性患者等相关。对这种家系的研究已鉴定了 *BRCA1* 和 *BRCA2* 基因。

一项对 1500 个家系的大规模分离分析（Newman *et al.*，1998）支持 4%～5%的乳腺癌，特别是早发病例可归因于遗传因素这一观点。收集呈近似-孟德尔遗传方式的家系，进行了连锁分析（图 15.7；有关连锁研究的细节见 MIM 113705）。1990 年，一个称作 *BRCA1* 的易感基因座定位于 17q21。与 17q-连锁的家系被诊断患病的平均年龄低于 45 岁。晚发家系的 lod 值为负值。随后在 1994 年，用 15 个与 17q 不连锁的乳腺癌大家系，鉴定了 *BRCA2* 基因座位于 13q12（MIM 600185）。鉴定这两个定位的基因的工作开始了激烈的“竞赛”。*BRCA1* 于 1994 年被克隆，*BRCA2* 于 1995 年被克隆——详细内容见 MIM 条目 113705 和 600185。*BRCA1* 可以解释 80%～90%既患有乳腺癌又有卵巢癌的家系，但只能解释少部分只患有乳腺癌的家系。男性乳腺癌患者主要发现于 *BRCA2* 家系中。

这两个基因都编码新的大分子蛋白质，这些蛋白最终成为转录的辅助激活因子，此外具有 DNA 修复作用（节 17.5.1）。它们作为肿瘤抑制基因而发挥作用，因为遗传的突变导致其功能丢失（节 16.3），家族性肿瘤丢失了野生型等位基因。然而乳腺癌的情形与结肠癌形成了鲜明的对比。在结肠癌中，通过对该病罕见孟德尔式研究所鉴定的 *APC* 基因，在常见散发型中也常存在突变（节 17.5.4）。相反，*BRCA1* 和 *BRCA2* 在散发型乳腺癌中很少失活。此外，这两个基因座的基因突变只能解释 20%～25%的患家族性乳腺癌的风险度（Pharoah *et al.*，2002）。一项对四个或以上乳腺癌患者的 257 个家系进行的调查（Ford *et al.*，1998）表明，在那些有四或五个女性乳腺癌而没有卵巢癌或男性乳腺癌的家系中，有 67%可能不涉及 *BRCA1/2* 突变。Nathanson 和 Weber（2001）进一步对易感基因搜索。分离分析的结果同易感性为多基因的观点是一致的，并已鉴定出一个常见的低外显率风险等位基因：在 1.1%正常

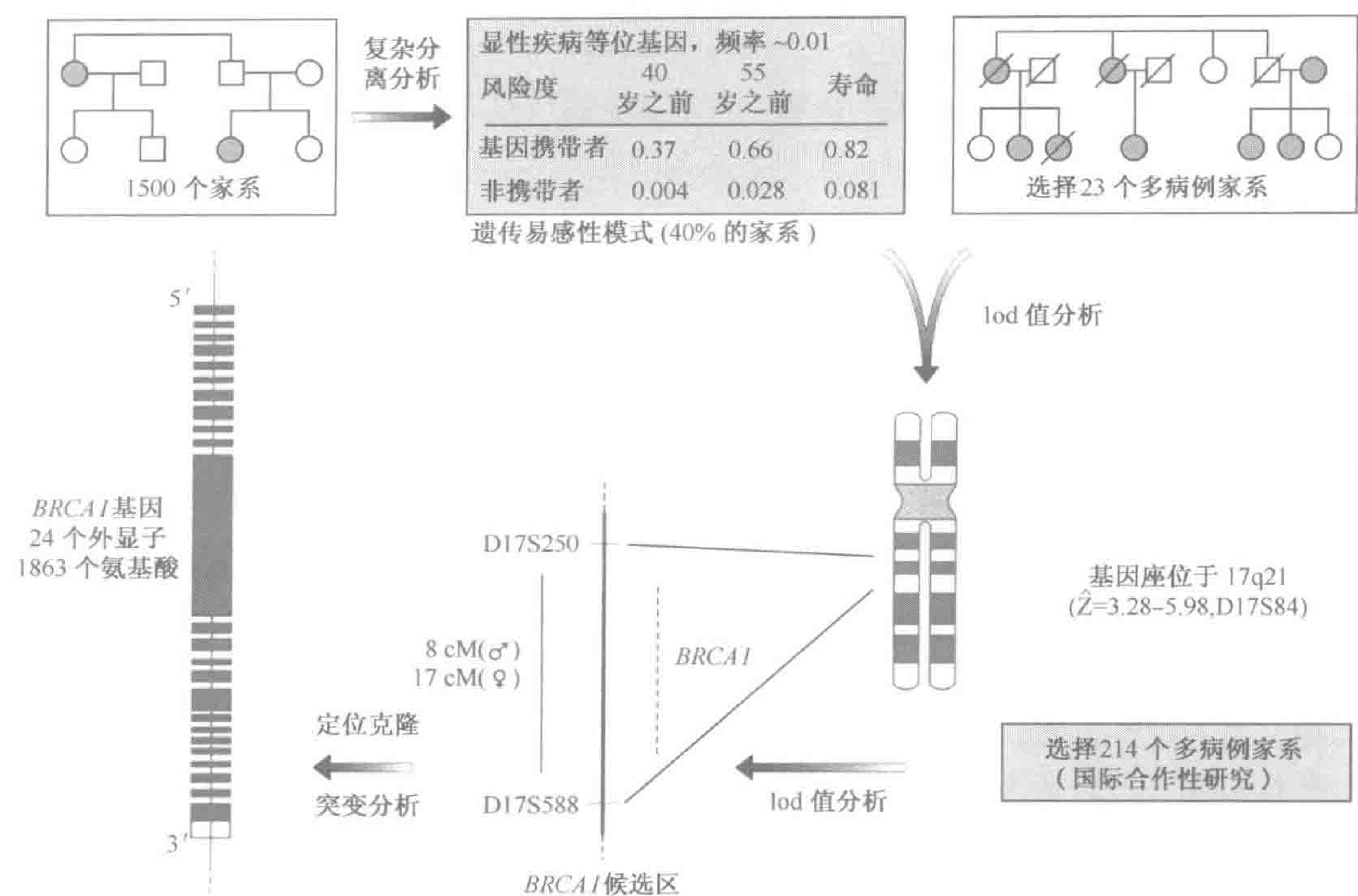


图 15.7　*BRCA1* 基因是如何被发现的

标准的定位克隆成功地鉴定了一种常见病的易感基因——但只是对于疾病的孟德尔遗传亚类。*BRC1* 看起来对常见、散发乳腺癌的作用很小。

人以及 5%乳腺癌患者中发现了 *CHEK2* 细胞周期激酶基因的突变（*CHEK2*-乳腺癌协会，2002）。

早期资料表明一个带有 *BRCA1* 基因突变的妇女有 85%～90%的可能性患乳腺癌，有 40%的风险度得卵巢癌。*BRCA2* 突变的妇女有相似的患乳腺癌的风险度（但不是卵巢癌）。然而，这个分析是在用于定位研究的大家系中进行。后来，在无家族史的妇女发现突变，她们的亲属中发现了更多非外显的病例。在北欧犹太妇女中，三种特定突变的频率相当高（*BRCA1* 的 185delAG 和 5382insC，*BRCA2* 的 6174delT）。对北欧妇女患乳腺癌且无家族史的亲属调查，显示这些突变携带者一生的患病风险度是 36%，而不是通常引述的 85%～90%（Fodor *et al*.,1998）。这是复杂疾病研究的一个普遍特征，即从家系收集最初鉴定易感基因座来初步估计外显率、严重性及亲属患病风险度，夸大了一般群体的风险度（Göring *et al*.,2001）。这一点对群体筛查有重要影响。最初家系研究鉴定的突变可能在筛查的病例中没有显著效应，另外也可能存在低外显率的突变在初始研究中被漏掉而在群体筛查时被发现的情况。

15.6.2　先天性巨结肠病：一种寡基因病

先天性巨结肠病（HSCR）是一种在部分或全部大肠或结肠中神经节的先天缺陷。缺乏蠕动的结果导致严重扩张的巨结肠，除非受累的节段被切除，否则这种巨结肠对新

生儿是致死性的。Amiel 和 Lyonnet (2001) 综述了 HSCR 的遗传学。18%的 HSCR 病例是一种综合征的症状之一，另有 12%的病例存在染色体畸变。剩下的 70%是单纯的 HSCR，一种典型的“多因子的”疾病，常常是家族性的而非孟德尔式的。我们已经看到了分离分析的结果（表 15.4）。对 HSCR 经典定位克隆的手段已经成功鉴定了一些基因座。

- **一种染色体畸变**：发现有两个 HSCR 的患者在 10q11q21 有可见的缺失。*RET* 癌基因位于缺失区，编码一种受体酪氨酸激酶，在相应的细胞中表达。研究结果表明，50%家族性和 15%～35%散发的单纯性 HSCR 患者存在 *RET* 突变，而一些未受累的亲属也存在这样的突变。已知的 *RET* 变异并不能解释 *RET* 基因座在易感性的全部效应；可能非编码变异也起一定作用（Gabriel *et al.*,2002a）。有关 *RET* 突变“有趣的”分子病理学见节 16.6.2。
- **连锁分析及小鼠模型**：通过一个多重近亲婚配 Mennonite 大家族另一个易感基因座定位到在染色体 13q；一些无亲缘 HSCR 患者的 13q 缺失也在该染色体区域。在一项研究内皮素控制血管张力的独立试验中，小鼠内皮素受体 B（*ednrb*）基因被敲除。出人意料的是，这种敲除的结果表现出能充分研究 HSCR 小鼠模型的表型：致死性花斑（s^l）。在人类，*ENDRB* 存在于 HSCR 的候选区域 13q，并且很快在 Mennonite 家族中证明了一个突变。有趣的是，一旦家系突变被确定（如 W276C），这一突变却显示出既不必需也不足以致病。外显率具有性别特异性且每一种基因型都不相同：对于基因型 W/W，外显率是 0.13（男）、0.09（女），对于基因型 W/C，外显率是 0.33（男）、0.08（女），对于基因型 C/C，外显率是 0.85（男）、0.60（女），这正表明了 HSCR 的寡基因性状。这项研究的报告正阐述了一种相对常见寡基因病基因鉴定的复杂性（Puffenberger *et al.*,1994a，b），很值得阅读。
- **直接检测候选基因**：既然 *RET* 和 *EDNRB* 都编码受体，编码它们配体的基因（*GDNF*，*NTN* 和 *EDN3*）自然就是候选易感基因。每一个基因的突变都证明存在于小部分家系中；另一个家系中还发现了 *ECE1* 基因的突变，该基因编码对内皮素 3 蛋白水解成熟所必需的一种酶。在另一个 HSCR 小鼠模型，显性巨结肠，一个候选基因 *SOX10* 被克隆鉴定，但后来证明在人类中，有 *SOX10* 突变的人群患有复杂的综合征，不是典型的 HSCR。位于染色体 2q22 的 *SMAD1P1* 基因突变在 HSCR 和智力低下的患者中很常见。
- **定位其余易感基因座**：*RET* 基因是一个主基因，但其外显率是不完全的且是性别依赖性的（男性 65%，女性 45%）。其他鉴定的基因偶见于家系中也很重要，但对整体易感性的作用不显著。Gabriel 等（2002a）实施了一项系统性的连锁分析，他们从中得出结论，即 *RET* 基因座的变异与 3p21 和 19q12 上的未知基因座变异以及 9q31 上的一个 *RET* 依赖性修饰基因一起相互作用可以解释 HSCR 的全部遗传易感性。然而，在上面提到的 Mennonite 家族中，易感性似乎是由 *RET* 和 *EDNRB* 基因座以及 16q23 一个未确定基因座的等位基因间相互作用决定的（Carrasquillo *et al.*,2002）。奇怪的是，既没有提示与 3p21 或 19q12 基因座关联，也没有提示与 21q22.3 基因座关联，而以前报道这些基因座在这一家族中有关联。

单纯性 HSCR 表现为一种寡基因病（Amiel and Lyonnet，2001）。如果定位克隆和

功能研究支持这一分析，HSCR 将成为这类疾病中第一个被充分剖析的疾病。对此成功的一个重要贡献是 λ_s 的极高值 187（Gabriel *et al*.,2002a）。

15.6.3　阿尔茨海默（Alzheimer）病：遗传因素在常见晚发型和罕见孟德尔遗传早发型中都很重要，但它们是不同基因以不同方式起作用

阿尔茨海默病（AD；MIM 104310）65 岁以上人群发病率约 5%，80 岁以上人群发病率约 20%。患者有进行性的记忆减退，伴随情感行为失衡和一般认知损害。脑部尸检揭示有神经元的丢失及许多淀粉蛋白斑片。变性的神经元含有特征性的神经元纤维结，罕见于早发年龄。早发和晚发 AD 的临床和病理特征是一样的，但有时早发疾病是孟德尔常染色体显性遗传，而晚发 AD 是非孟德尔式，只显示中度的家族聚集性。显性早发家系的 lod 值分析定位并随后克隆了 3 个基因，21q21 的 *APP*、14q24 的早老素-1 和 1q42 的早老素-2（分别见 MIM 104300，104311 和 600579）。尽管这三个基因座的突变只能解释 10%的 65 岁以前发病的病例，它们尤其见于罕见的外显率极高且发病极早的显性 AD 家系中。如果任何人想知道这种疾病对一个家庭意味着什么，那么就应该阅读 Daniel Pollen 的《Hannah 的继承人》(进一步阅读)。

多病例的晚发家系没有证据显示与早发涉及的基因座连锁，但显示了与 19 号染色体连锁。最后，这个易感基因座鉴定为 *APOE*（载脂蛋白 E，MIM 107741），定位于 19q13.2。该基因座有 3 种等位基因，其频率（高加索人中）为 0.08(E2)、0.77(E3) 和 0.15(E4)。家族性和散发性晚发 AD 都与 E4 等位基因强烈关联，而 E2 与抗 AD 关联。对 65 岁以上的一项横断面研究中，与 E3 纯合子相比，E3/E4 个体约有 3 倍的患病风险度，E4/E4 个体则约 14 倍。ApoE 似乎能解释约 50%的晚发 AD 易感性。因此，与乳腺癌形成对比，不单罕见的孟德尔遗传型，而且常见的散发型也有较大程度的遗传决定性。一项纵向研究（Meyer *et al*.,1998）表明，E4 可能控制发病年龄而不是易感性。一旦 AD 发病，其进展对于 E4 和非 E4 的个体没有不同。尽管人们对 E4 与晚发 AD 的关联没有争议，其机制却不清楚。E2/E3/E4 决定因素是氨基酸替换 R112C 和 R158C；深入搜查没有显示任何与 E4 连锁不平衡的"真正的"易感性决定因素。

进一步的连锁和关联研究已产生了过多的其他易感候选区域。Emahazion 等（2001）列出了这些区域，并报告了用 60 个候选区域或基因的 SNP 进行的一项大规模的关联研究。原始 p 值经多重检测校正后，研究结果只是鉴定了 *APOE*，而其他所有候选都是阴性。作者用这些数据只是强调：确认或驳倒已宣称的复杂疾病关联分析是多么的难。追踪研究通常无法重复最初的定位——有多大概率是因为最初的报告犯了Ⅰ类错误？又有多大概率是因为追踪研究缺乏效能？由于研究的统计学倾向于过高估计他们鉴定的真正基因座的效应（Göring *et al*.,2001），所以很难确定追踪研究要驳倒一个已宣称的连锁需要多大的效能。

ApoE 的情形与常见病易感等位基因鉴定中涉及可能的社会和伦理问题讨论也很相关（伦理学框 1）。

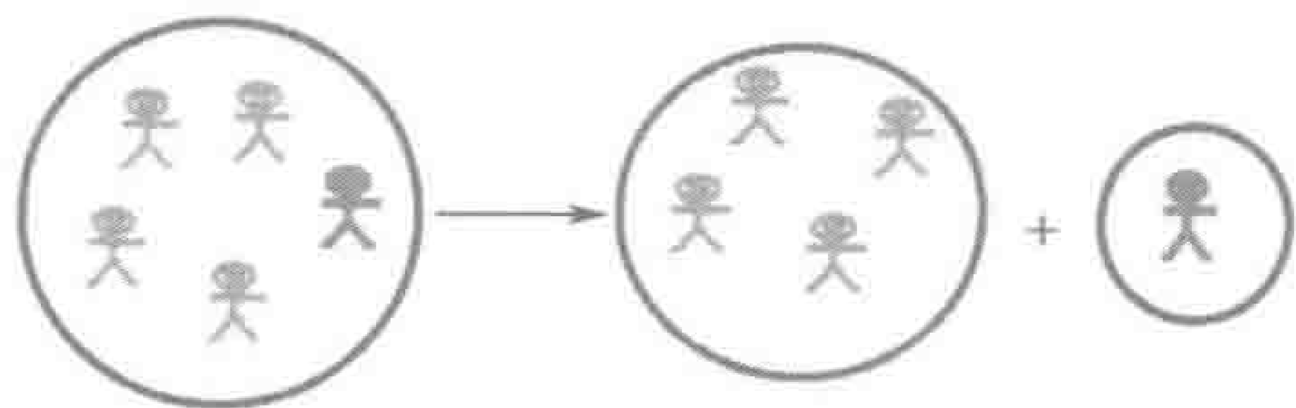
伦理学框 1　阿尔茨海默病、ApoE 检测与歧视

对常见病易感的基因型鉴定引发了伦理和社会问题，ApoE 就是个很好的例子。雇主和保险公司会力图广泛应用这些检测吗？这会导致就业、医疗保健和保险方面的不公平歧视吗？

就业。雇主很合理地对某人做这份工作的现有能力感兴趣，而对十年内会发生什么情况不太感兴趣。需要雇主大量投资进行专业培训的工作在这方面会有部分例外。在英国，雇主对现在健康的员工因其以后有患某病的风险而歧视是非法的。

医疗保健。在医疗体系社会化的国家，医疗保健的歧视问题不会发生。在哪里可能涉及取决于私人保险，问题又回到了保险方面的歧视。

保险。商业保险以相互关系的原则为依据。担保人将申请人划分到危险群体，保险费会反映出申请人被引入该群体的风险。每一项经营的保险单都存在保险精算的公正与道德的公正之间的冲突。一个申请者因为自己所不能控制的基因结构而受到惩罚，这公平吗？然而大部分人对保险公司在男性和女性收费率不同没有疑问。明确的解决方案是确保基于相互关系的保险不用于提供文明生活的必需品，如基本卫生保健，而要将其被限定为一种产品，在消费者社会作为人们正常选择的一部分，可以选择买还是不买。



风险度群 A　保险费£1.00　　风险度群 B　保险费£0.05　　风险度群 C　保险费£3.00

与普遍观念相反，保险公司并不感兴趣于让人们参加基因检测。在上面阐述的例子中，利用基因检测将危险群体 A 分为危险群体 B 和 C 对保险公司并没有好处。保险公司的兴趣是避免逆选择。当申请人利用秘密了解来获得不公平的利益时，就会出现这类问题。私下知道自己有很高患病风险的一些人按照标准价钱买特别大的保险单，而知道自己有很低患病风险的一些人决定不买保险。因此保险公司想知道相关先于检测的信息是同申请人一样多；让申请人做额外检查的唯一合理的理由是，他们怀疑申请人实际上已秘密地做了。

什么是“相关”的标准对于临床检查和保险检查是不同的。对于有用的临床检查，其结果对每个个体都必须是预言性的，对于保险检查，理论上它只需要能够区分出具有显著不同的平均患病风险度的群体就行——一项低要求的检测。这就是平均标准差和标准误的不同。*ApoE4* 既不必需也不足以导致 AD。有 *ApoE4* 的人约一半永远不会患 AD，不管他们活多久，而很多 AD 患者没有 *ApoE4*。这种不确定性使它既不能确认临床诊断，也不是对健康人的一项有用的预测性的检测。美国医学遗传学学院和美国人类遗传学学会曾建议，ApoE 检测不应该用于 Alzheimer 病常规的临床诊断或预测性的检测（ACMG/ASHG 工作组，1995）。对 Alzheimer 病，ApoE 基因分型的唯一临床指征是针对治疗 Alzheimer 病患者的昂贵药物，如果只对某种 ApoE 基因型有效，或者尤其是如果对某种基因型有害。对于保险，关键问题是 ApoE 基因型能否确定的风险度完全不同的群体，从而使这项检测有意义。据有资质的保险精算师 Macdonald 和 Pritchard（2001），基本回答是“否”。

尽管 ApoE 检测没有引发重大的公共政策问题，其他检查可能会出现而需要社会控制。一项敲响警钟的检查有三个特征：

- 它应该能预测患病风险，该风险在生活方式和家族史中并非明显（这除外，例如，Huntington 病检测）；
- 疾病应相当常见，且检测有足够的预测性，使得保险公司不能冒险忽略它，但该检测应该不能作为常规临床服务的一部分；
- 可信的私人检测允许获得并被广泛使用。

15.6.4　1 型糖尿病：依然是遗传学家的梦魇？

迈向解析糖尿病遗传学的第一步是区分糖尿病的不同类型（表 15.8）。1 型和 2 型是不同的疾病，有不同的病因、不同的自然病程和不同的遗传学。1 型糖尿病（T1D，胰岛素依赖型糖尿病；MIM 222100）是由胰腺 β 细胞的自身免疫性破坏引起，典型受累年轻人需要终生胰岛素治疗。本病有相当强的家族聚集性（$\lambda_s=15$，MZ 双生子一致性大约 30%）。于 20 世纪 70 年代已确定了与某些 HLA 等位基因的关联。

表 15.8　糖尿病的临床分型

1 型糖尿病	2 型糖尿病	MODY
青少年发病	成熟期发病（> 40 岁）	青少年发病
占英国人群的 0.4%	占美国人群的 6%	罕见
需要胰岛素	通常由口服降糖药控制	同 2 型糖尿病
无肥胖症	与肥胖症强烈关联	无肥胖症
家族性：	家族性：	家族性：
MZ 双生子一致性 30%	MZ 双生子一致性 40%～100%	常染色体显性？
同胞患病风险度 6%～10%	同胞患病风险度 30%（可能为亚临床的）	
与 HLA-DR3 和 DR4 关联	无 HLA 关联	无 HLA 关联

MODY（青年人中的成年发病型糖尿病）是一种不常见的孟德尔遗传式糖尿病，已经对它鉴定了各种基因（MIM 600496）。另外，糖尿病可以是许多罕见综合征的一部分。

在英国，约有 95%的 T1D 病人有 HLA-DR3 和/或 DR4 抗原，相比之下，一般群体只有 45%～54%。虽然 HLA 可以解释约 40%的遗传诱因，但某些不同单体型与易感性或抵抗性关联。位于主要组织相容性复合物内的强烈连锁不平衡使得鉴定易感性的原发决定因素很难。已证明，与糖尿病的低风险度关联的单体型均在 *DQB1* 基因座带一个等位基因，该位点氨基酸 57 为天门冬氨酸（Asp），而高风险度单体型 *DQB1* 等位基因带有其他氨基酸。在一项研究中，有 96%的糖尿病患者在 *DQB1* 57 位点是非天门冬氨酸的纯合子，而对照人群只有 20%（Todd *et al.*,1987）。可能一个抗性因素是由共享 Asp-57 的几种不同 HLA 分子抗原决定簇确定的。这种共享抗原决定簇的观点也已运用到其他 HLA 相关的自身免疫病，并取得了不同程度的成功。

T1D 的早期研究也鉴定了另一个易感基因座 *IDDM2*，离胰岛素的结构基因 *INS* 很近。连锁不平衡定位已表明，真正的决定因素是 *INS* 基因上游的一个 14bp 微卫星重复。易感性与短重复（26～63 个重复单位）相关。长重复（140～210 个重复单位）导致 *INS* 基因在发育胸腺中呈相当高水平的表达；据称，在免疫系统发育过程中，这会增加胰岛素反应性 T 细胞克隆的丢失的效率，因此减低了自身免疫性攻击的风险度。*HLA-DQB* 研究和 *INS* 调查结果强化了导致常见病易感性的微小遗传改变和见于孟德尔遗传病的重大改变这两种类型间的可能区别。

一些研究小组已积极推崇全基因组连锁扫描策略后的候选区域内连锁不平衡定位。欧洲协议（2001）报告给出了一些重要连锁研究的参考资料，表 15.9 列出了迄今为止涉及的主要区域。为了防止强烈的 HLA 效应掩盖微弱的信号，数据通常按照 HLA 基

因分型进行其他基因座的连锁分析。由 ASP 研究确定的区域很广，另外“真正”的基因座也可能远离 lod 峰值的地方，这使我们很难知道比如 2q 和 6q 有多少不同的易感基因座。理论上讲，每一个基因座的全部信息可以通过 λ_s 表示为 T1D 所有 λ_s 的分数来衡量（Risch，1990a），以便使我们知道还剩多少基因座要阐明。然而，通常做法是用同一个数据组先确定一个候选区域，然后估计 λ_s，这会得出非常不可靠的估计（Göring，2001）。HLA 和 *INS* 大概解释 50%的易感性，但我们并不知道表 15.9 中其他基因座的全部信息。

表 15.9　由受累同胞对（ASP）和传递不平衡（TDT）分析证实的 1 型糖尿病的主要易感基因座

基因座	MIM 号	位置	地位
IDDM1	222100	6p21	$\lambda_s=3.1$；决定因素是 *HLA-DQB*
IDDM2	125852	11p15	$\lambda_s=1.3$；决定因素是 *INS* 基因上游的一个 VNTR
IDDM4	600319	11q13	$\lambda_s=1.6$；三次筛查(596 个家系)综合的结果中显著连锁
IDDM5	600320	6q24-q27	$\lambda_s=1.2$；四项研究中观察所得
IDDM6	601941	18q21	一项研究中的 ASP 和 TDT 证据
IDDM7	600321	2q31-q33	$\lambda_s=1.3$；三项 ASP 研究所见，与候选基因 *CTLA4* 的连锁不平衡
IDDM12	600388	2q33	$p=5\times10^{-5}$但只见于一些群体
IDDM8	600883	6q25-q27	$\lambda_s=1.8$；与 *IDDM5* 没有明确的不同
IDDM10	601942	10p11-q11	三项研究的 ASP 和 TDT 数据
IDDM13	601318	2q34	与 *IDDM7* 和/或 12 相同?
IDDM15	601666	6q21	确定的(尽管很难与 HLA 效应区分)

数据来自 OMIM 条目及其中的论文；λ_s 值来自 Luo 等(1995)。

T1D 的一些特征似乎很支持遗传学研究。它相当常见（每 1000 人中有 3 个），以便可以收集大量患者群体来研究。患者年轻而其双亲通常适合 TDT 分析。诊断明确，且有很好的动物模型，以 NOD（非肥胖糖尿病）小鼠为模型中的疾病也是多基因的。由于所有这些原因，并且因为其巨大的财政上和人力上的花费，数十年来一直对 T1D 进行了连锁分析以及候选基因和模型生物的研究。由于迄今只有相对微小的进步，可再次使用那句对糖尿病的古老描绘，即遗传学家的梦魇。

15.6.5　2 型糖尿病：两个易感因素，一个常见且连锁检测不到，另一个很复杂且只存在于某些群体

2 型或非胰岛素依赖型糖尿病（T2D）是这一异质性疾病最常见的形式，全世界大约有135 000 000人受累。T2D 是胰岛素分泌受损和终末器官反应性降低的联合作用的结果；已知的风险因素包括年龄、肥胖和缺乏运动。在许多发达国家，10%～20%的 45 岁以上人口患病，某些群体的发病率特别高。

已报道，T2D 与至少 16 个不同的遗传变异关联（Altshuler *et al.*，2002a 总结）。Altshuler 的大样本研究只重复出所有这些遗传变异中的一个。在某些群体中 *PPARG* 基因的一个常见等位基因（$p=0.15$）与降低 T2D 患病风险度关联（$OR=0.8$）。由于增高患病风险度的等位基因很常见（$p=0.85$），连锁分析不可能检测到。*PPARG* 并非

一个令人惊奇的候选基因：它编码一种核激素受体，该受体能调节脂肪形成，并且是用于治疗 T2D 的噻唑烷二酮类药物的靶点。

Altmüller 等（2001）详述了 10 项全基因组扫描，报道了 T2D 与 15 个不同染色体上 25 个基因座间的强烈连锁。像通常那样，各项研究间不能很好地重复这些结果。在墨西哥裔美国人中鉴定了一个基因座，即位于 2q37 的 *NIDDM1*，它与 15 号染色体上一个基因座的相互作用增加了这组群体的易感性（参考资料见 Horikawa *et al*.,2000）。Horikawa 等证明了 2q37 的钙蛋白酶 10（calpain 10，*CAPN10*）基因内的一个 SNP 特定组合是墨西哥裔美国人的易感性决定因素。引述一篇很值得阅读的社论，"这项研究在复杂性状定位克隆的关键时刻展现了一个崭新的景象，并告诫关于发现常见病基因确定多么困难的持久重要性。"（Altshuler *et al*.,2000b）。

这项研究开始是将候选区域缩窄到 2q37 的一个 7cM 区域。这在物理位置上相当于一个 1.7Mb 的重叠群（越靠近端粒重组率倾向于更高，节 13.1.5）。检测了该区域一系列多态性，不仅为了与 T2D 的关联，也为了与连锁证据的关联。这样做的基本理论依据是病例中只有一种亚类与 2q37 基因座连锁，而应该是携带易感性决定因素的个体。最初的研究提示了一个 66kb 的靶区，经检测发现该区包括 3 个基因，*CAPN10*、*RN-PEPL1* 和 *GPR35*，以及 179 个序列变异（在 10 个墨西哥糖尿病患者上）。这引起了对变异的鉴定，UCSNP-43，该处 G/G 基因型纯合子显示与连锁证据相关联，而且也可能与糖尿病关联（OR1.54；可信区间 0.88～2.41）。后来开始了单体型的搜索，这些单体型：(a) 在 2q37 最大连锁的患者组中频率增高；(b) 受累同胞对共享的频率比预

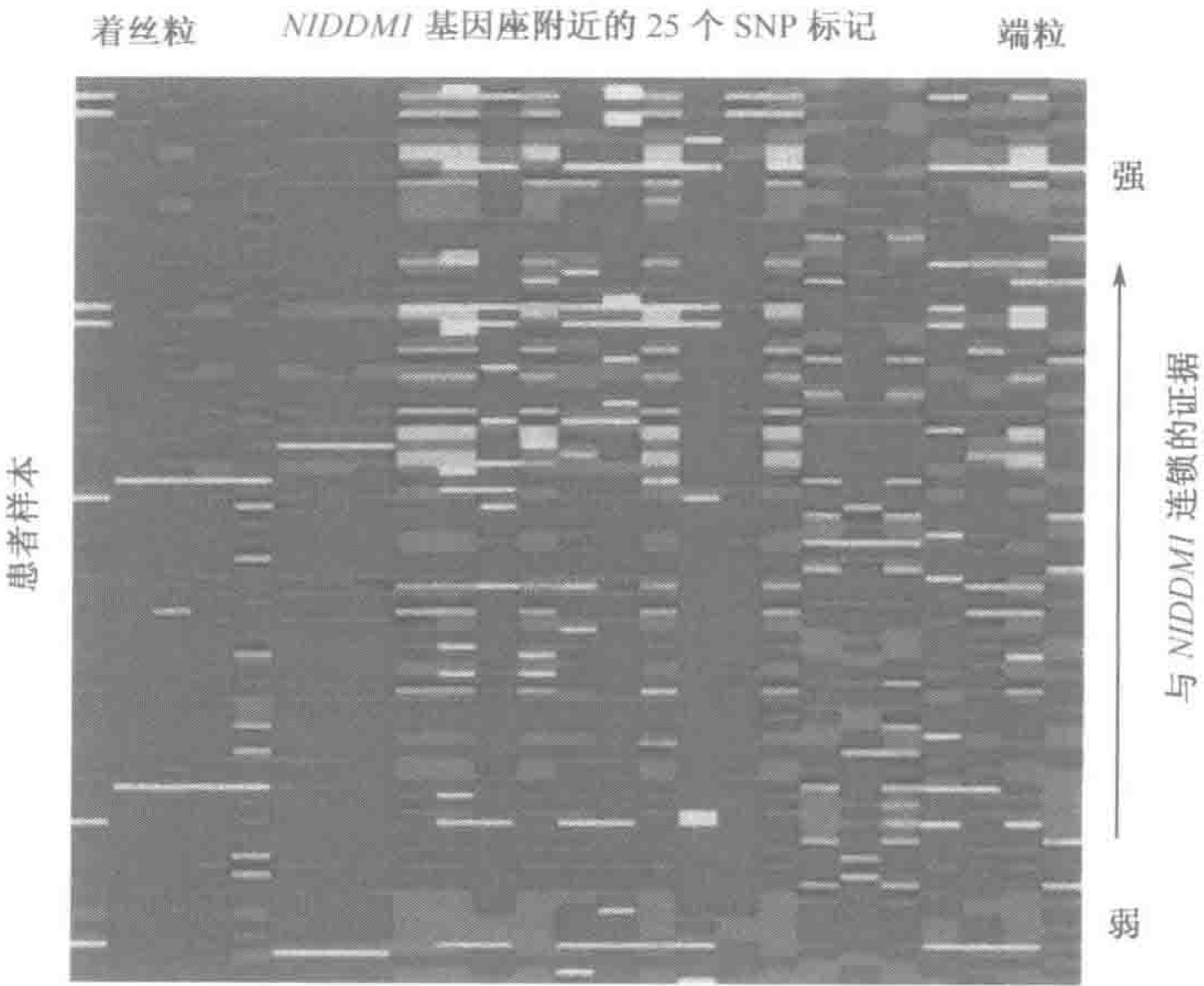


图 15.8　2 型糖尿病患者在钙蛋白酶-10 基因座的基因型

每一行总结了一个患者 25 个 SNP 的结果（列）。蓝色：常见等位基因纯合的；红色：杂合的；黄色：罕见等位基因纯合的；白色：无数据。患者按照与 *NIDDM1* 连锁证据逐渐降低的顺序从上到下排列。注意图中间部分，*CAPN10* 基因的位置，其颜色从上到下是怎么变化的，而其他部分的颜色不这样变化。这显示了位于 *CAPN10* 基因座的某种基因型与连锁证据关联的程度。经牛津大学出版社允许，由 Cox N J（2001）Hum. Mol. Genet. 10，2301～2305 改编而来。

期的更高；并且（c）与增高的糖尿病风险度关联。两种单体型的杂合型（由 *CAPN10* 基因内含子的 3 个 SNP 确定）满足了所有的标准。这两种单体型的纯合型都不是危险因素。因此 *NIDDN1* 的易感性决定因素似乎是非编码 SNP 的一种特殊杂合组合与 15 号染色体上一个未鉴定因素之间的相互作用。图 15.8 试图给我们“与连锁证据相关联”的意义的视觉印象。

怎样才能坚信这次轰动性的项目已真正鉴定了 *NIDDM1* 易感因素？后续的研究已得出了矛盾的结果（Fullerton *et al*.,2002 总结），这一易感性基因型与 T2D 的关联在全球范围内确实没有一致。*CAPN10* 是一个出乎意料的候选基因；而 UCSNP-43 对基因转录效应已有报道，并提示了一种可能是下游的胰岛素刺激葡萄糖转换作用。总之，这项研究促使我们在鉴定复杂疾病易感因素方面相当谨慎。从正面看，一个新的候选基因已被鉴定；从负面看，我们还远不清楚导致易感性的真正变异是否已被鉴定，而且也不清楚这种不确定性该怎么解决。其他常见病的危险因素也会同样地含糊和复杂吗？它们在特定群体会有特异性吗？Horikawa 等不懈的努力至少提供了许多精神食粮。

15.6.6 炎症性肠病：一个明确易感性基因的鉴定

溃疡性结肠炎（UC）和 Crohn's 病（CD）是炎症性肠病（IBD）的形式，在西方世界，每 1000 人中有 1～3 人受累。家系和双生子研究在某些共享因素上支持疾病的遗传易感性。λ_s 对于 CD 为 20～30，对于 UC 为 8～15。许多研究小组通过同胞对和家系连锁分析鉴定了一些已确认的或可重复的连锁区域（*IBD1*，16p12-q13；*IBD2* 12p13；*IBD3* 6p；*IBD4* 14q11-q12；*IBD5* 5q31）和提示连锁的其他许多区域（参考文献见 OMIM 条目 266600）。特别是通过一项 613 个家系 1298 个 CD 和/或 UC 受累同胞大样本研究，以最大 lod 值 5.79 确定了与 *IBD1* 的连锁（IBD 国际遗传学协作组，2001）。

最近，三个研究小组鉴定了在 *IBD1* 区域的基因（Hugot *et al*.,2001；Ogura *et al*.,2001；Hampe *et al*.,2001）。Ogura 等克隆了 *CARD15*（*NOD2*）基因，它是一个 NF-κB 炎性反应的调节因子。他们在 IBD 患者中检测该基因，因为它定位在 *IBD1* 基因座而且生物学上可能是一个 IBD 候选基因。相反，Hugot 等采用的是纯连锁-关联分析途径。通过连锁缩窄候选区，而后着手做关联研究。有趣的是，对 *CARD15* 进行的 TDT 结果只是临界显著（$p=0.05$），尽管该基因在一个独立的样本中得到了重复 $p<0.01$，但这一次却是另一个不同的标记！实际上，仔细观察他们的数据发现，在一个大群体中，随机的基于微卫星检索的关联不会有阳性结果。最终，尽管不引人注目但强烈的关联强调了一个基因组区域，他们从该区克隆出当时尚未发表的 *CARD15* 基因。

IBD1 基因座的首要易感因素是 *CARD15* 基因 C 端的单个碱基插入 3020insC。Hampe 等证实了 CD 和这一突变之间非常强烈的 TDT 关联。这一插入在 CD 患者中的频率为 0.08，对照中的频率为 0.02，而在 UC 患者中的频率为 0.02，说明这一变异是 CD 的易感因素而不是 UC 的易感因素。杂合子患 CD 的相对风险度是 3，而纯合子患 CD 的相对风险度是 23。这一插入导致了一种蛋白质 33 个氨基酸的截短，影响了一个对 NF-κB 激活至关重要的区域。

这一研究强调了三点。首先，它表明这样的策略是可行的：易感基因至少有时候能够通过遗传学途径鉴定。其次，它强调了对大规模家系研究的需求。最后，Hugot 等研究中的连锁标记微弱 TDT 结果与 Hampe 等直接检测插入突变得出的强烈结果进行比较，指出了关联研究的重要意义。*CARD15* 基因的其他突变也导致了 CD 的易感性这一点并不奇怪。两个其他的 cSNP，G908R 和 R702W，（Hugot 等另有不同的编号），与 CD 强烈关联，且 *CRAD15* 基因内的多种多样的罕见错义突变整体上在 CD 患者中比对照更常见。尽管存在连锁不平衡，每种突变都会与它各自的单体型关联。这种等位基因异质性完全在意料之中——但严重限制了在候选区域采用以关联为基础的策略定位易感基因的机会。

15.6.7　精神分裂症：精神病或行为障碍的特殊问题

研究精神疾病或反社会行为的遗传学家面临两个额外的问题，这两个问题是研究复杂疾病所有困难之首。

（Ⅰ）本性-养育争论。由于共享的家庭环境对于精神或行为问题在家系中有世代遗传的倾向是一个合理的解释，声称遗传因素起作用要有一个高标准的证据。对于精神分裂症，表 15.1～15.3 总结了从家系、双生子和领养子研究中得出的一些证据。大多情况下，本性-养育争论只是政治争论的替代品，在街头政治（street politics）的体面（decorum）和客观上运作。争论双方把他们的热情置于一种错误的前提下，即如果一种疾病是遗传性的，就无法被社会干预所改善。左翼人士支持环境原因，右翼人士支持遗传原因。目前一些热点已不再是有关精神病学的争论，但研究反社会行为的遗传学家依然引来了恶毒的反抗和不欢迎的支持者。

（Ⅱ）诊断标准。正如我们在节 15.3.1 中讨论的，诊断标准是精神病遗传学的另一个非常困难的地方。“精神分裂症不应该被看作一个疾病，而应该像癫痫一样被看作是一个综合征，由一组体征和症状来识别，且有不同的发病原因”（Trimble，1996）。比如 DSM-Ⅳ的诊断，达成一致的标准是依据有经验的精神病学家对组成该病的重要核心所作的最佳判断——但根本上说它们是主观的。也许“精神分裂样个性”是使人们易感 DSM-Ⅳ精神分裂症的基因的表现之一？或者，也许对精神病存在一个普遍的遗传易感性，躁狂和抑郁症也应包括在内？遗传学分析常用两套或更多不同的诊断标准，一套范围狭义而另一套范围广泛，来看看哪一套得出最好的 lod 值。此外，研究者可以寻找与中间表型的连锁，中间表型不是精神分裂症，但可能形成易感性的一部分且可能更单纯地由遗传决定——可能是一些生理的或神经心理的变异。

精神分裂症作为一种相对常见又极其令人痛苦的家族性疾病，长期以来一直高居遗传学研究项目的榜首，Altmüller 等（2001）列出了自 1994 年以来实施的 10 项全基因组扫描，还有许多对个别候选基因或候选区域的其他研究。典型的候选基因编码参与神经传递的蛋白或受精神分裂症治疗药物影响的蛋白。多发受累的大家族并不罕见，这些家族经常被用于连锁研究。选择这样的家系意味着采用“Genehunter”分析：Genehunter 是复杂疾病连锁最常用的程序，不能处理大的系谱（节 13.6.2），使得许多研究人员采用参数 lod 值分析（节 13.3.1），但参数 lod 值分析有各种可选择的遗传模式。虽然这样做很有效，但又因引入额外的自由度而降低了研究的效能。研究方法与多种诊

断标准的使用结合（见上文），使评判结果的显著性变得异常错综复杂。然而，生物学上的多重模式是现实的——在某一基因座上的一个危险等位基因可以决定泛指精神分裂症的显性易感性，而另一基因座的一个等位基因可以决定狭义精神分裂症的隐性易感性。

目前一项或更多研究显示了十余个染色体区段连锁的证据，但没有一个易感等位基因得到清楚地鉴定。一个有可能的候选基因是 *COMT*（儿茶酚-*O*-甲基转移酶）的 Met158Val 多态缬氨酸等位基因。生物学上这是一个可能的候选，尽管有些研究是否定的，但大量研究已报道了肯定的关联和 TDT 结果。例如，Weinberger 等（2001）发现，该基因在对照中频率为 52%，而在精神分裂症患者中频率为 60%，纯合子的 *OR* 值为 1.5。然而，尽管已经得到确认，这个因子只能解释风险因素中 4% 的变异。OMIM 181500 提供了主要候选区域相关论述的链接，Gurling 等（2001）、Camp 等（2001）、Weinberger 等（2001）和 Lewis 等（2003）的研究是这些问题和某些方法的很好展示，一些方法正用来试图解释一残酷疾病的遗传学研究。这里出现的问题同样适用于各种其他的人类精神病和行为表型。

15.6.8　肥胖症：数量性状的遗传学分析

作为一个有争议性的情感问题，肥胖症与精神分裂症相匹敌，瘦人中盛行的观点将肥胖症归咎于自制性差和意志松懈，肥胖派则归咎于身体的素质。肥胖症是发达国家公共卫生关注的主要问题，也引起了生物技术公司的极大兴趣，他们发现了一种有效的减肥药所能带来的财富。就我们现在的目的，对肥胖症主要的兴趣在于将其作为数量性状的一个范例（Barsh *et al*.,2000）。

单纯测量体重并不能为分析提供一个很好的数量变量，因为这会把高瘦的人和矮胖的人混在一起。体重指数（BMI；体重 kg/身高 m^2）是一个更好的衡量指标。在美国，1983 年平均 BMI 为 $22kg/m^2$。其他研究者已尝试采用更接近基础生理学的衡量指标，例如，脂肪组织百分率或血清瘦素（leptin）浓度。每一种衡量指标都可以用某些分值来定义肥胖症，例如，经按照年龄和性别调整，BMI25～$29.9kg/m^2$＝Ⅰ级，BMI30～$40kg/m^2$＝Ⅱ级，BMI＞$40kg/m^2$＝Ⅲ级。然而，这样一种方法是武断的，并且会丢失数据。直接分析数量变量来鉴定**数量性状基因座**（quantitative trait locus，QTL）会更好。这些数量衡量指标是动植物生长的根本。人类家谱的 QTL 分析常用技术是变异-组分连锁分析。Almasy 和 Blangero（1998）描述了这种方法的数学基础。本质上，正如传统的 lod 值分析，从两种不同假设的数据似然性比值算出一个 lod 值。这种情况下得出的数据是数量表型亲属间的协方差，不同的假设是染色体位置存在还是不存在这个 QTL。协方差可以按年龄和性别之类的混杂变量进行调整。按照 QTL 存在的假设，如标记数据所显示的，其效应可以从亲属对的染色体片段的遗传一致性预测到。Almasy 和 Blangero 描述的多重方法提供了 QTL 位置的可信区间和对其效应的一个估计。尽管我们将 BMI 视为一个 QTL 而不是将肥胖症视为双歧性状来分析，还是关注两个极端分布的群体达到最大的统计效能。

OMIM 在标题或临床摘要中有 44 个“肥胖症”的条目，包括罕见的严重肥胖症，这些肥胖症涉及调节食物摄取途径的组分——瘦素激素、瘦素受体、阿片-促黑素细胞

皮质素原及黑皮质素 4-受体。然而，大多数肥胖症是非孟德尔遗传性的。实际上，伴有不可避免的肥胖综合征并不引起人们的兴趣：对肥胖症研究的目的是发现易感或抵抗饮食诱发肥胖的机制，而不是控制体重本身的机制——对动物模型，这种观点同样适用。这种想法提示，最佳的研究设计应该考虑的只是那些肥胖者或不胖者以垃圾食品为生且不运动的。他们大都可以通过免下车的快餐出口处获得食品。

像平时一样，第一个任务是检查病因中是否有遗传因素。肥胖症倾向于在家系中世代相传，但这很容易是共享家庭环境的结果。然而，双生子和领养子研究表明约 70％的 BMI 变异可归因于遗传因素。分离分析提示，20％～65％之间的 BMI 群体变异可能是由于 1～2 个隐形主基因的作用（Feitosa *et al.*,2002 总结）。Altmüller 等（2001）列出了 2000 年 12 月以前报道的五项全基因组扫描；Feitosa 等（2002）和 Deng 等（2002）进行了两项更新的报道。后两项报道都显示了每一条染色体的 lod 值曲线，这些曲线很清楚地阐明了无法重复实验结果这一普遍性问题。就像其他复杂疾病一样，关键问题是，无法重复告诉我们，初始的报道是错误的还是后继的研究缺乏效能。在肥胖症的事例中，还值得我们反省的是，一些最好的数据可能藏于生物技术公司文件中，直到获得了一些专利才会展现。

15.7　概要及总结

15.7.1　为什么会如此困难?

鉴定易感因素证明比 10 年前大多数人想像的要困难得多。Weiss 和 Terwilliger（2000）恰如其分地提出："在申请资助提案所承诺 80％效能的复杂疾病基因定位项目中，有多少百分比是真正成功地鉴定假设的易感基因?" 迄今为止，答案是低于 80％，肯定是出了什么问题。任何觉得复杂疾病研究会很容易的人都应该阅读他们的论文。然而，未知领域的每一次大探索总是有激烈的学术争议伴随，其观点是为什么永远成功——然后又被证明是错的。大多数人的观点其失败是由于缺乏效能，这可以通过更大的群体，更精确的基因分型和一个更有力的统计学工具得到补救。

核心问题肯定是异质性

受累同胞对分析是连锁的一个非常健全的方法。得到如此小的显著 lod 值且各项研究间互相重复如此罕见（Altmüller *et al.*,2001）的事实只是由于基因座异质性。很明显，大多数复杂疾病易感性是由许多微效基因座而不是由少数主要基因座决定的。同胞对分析受到检测相当强的易感因素的限制。表 15.7 中的计算结果表明，对于 $\gamma \leqslant 2$ 受累同胞对共享的等位基因很少，检测它需要很大的样本数。Altmüller 等（2001）比较了 101 项全基因组扫描，想了解是否能得到成功的经验，但除了大样本分析这样明显的事实外，没有鉴定出任何"金"规律。另外要考虑的是，不同群体的易感性可能有不同的遗传学原因。例如，对日本 483 名 Crohn 病患者的一项调查（Yamazaki *et al.*,2002）没有发现 *CARD15* 突变的证据，而 *CARD15* 是欧洲人的一个主要易感基因（节 15.6.6）。

关联研究是另一个主要工具，依赖于连锁不平衡。有一点已经越来越清楚，当我们

有了图 15.6 很好的不平衡线性图时，才有选择哪个标记是合理的依据。然而，现在也很清楚的是，与 Kruglyak（1999）的计算结果相反，不平衡在很多病例中足够广泛，以目前的研究规模，足以给 SNP 关联研究合理的成功机会。既然已选择了合适的 SNP 来检测，主要问题就在于等位基因异质性上。炎症性肠病（节 15.6.6）提供了一个完善的解释方式，即一旦危险基因座出现不止一个常见易感等位基因，关联研究的效能就会迅速骤降。Risch 和 Merikangas（1996）的计算对推进 TDT 到复杂疾病研究前沿是有影响力的，但是他们只考虑了疾病基因座的单一等位基因。他们的计算说明那个等位基因，而不是基因座对该疾病的所有影响，记住这一点很重要。易感基因座等位基因异质性的程度成为成功或失败的决定因素。大多数易感等位基因是古老常见的多态性（“常见疾病-常见变异”假说），还是罕见新近突变的异质性汇集，就像大多数孟德尔遗传疾病？双方已产生了争论（Reich and Lander，2001；Pritchard，2001；Wright *et al.*,2003）。裁决仍在进行中。

15.7.2 如果所有问题都解决了，而且我们鉴定了易感等位基因——然后呢？

假定鉴定疾病易感因素获得成功，一定会提高我们对疾病发病机制的理解。但不会自动地改进治疗——尽管我们很好地理解了发病机制，该病可能仍然是不可治的，有时候纯粹的症状治疗非常有效——但是了解病理学应该使治疗更合理性和靶向。不同基因型的患者可能会有不同的潜在病理学，因此对不同的药物有不同反应。

鉴定危险因素是否会导致有效的预防尚不清楚。高通量基因分型的技术发展将使群体筛查更容易，但始终要问这种筛查是否合算，在伦理上是否可以接受。必然条件在节 18.3 中已有详细的讨论。最重要的一点，比任何技术问题更重要，就是鉴定某人有患病风险应该采取一些有用的行动。一般是谈论生活方式的改变。预防性的药物治疗只有针对少数高危群体，而且这些人有孟德尔式（Mendelian）特征而不是复杂疾病时，才是无可非议的。Pharoah 等（2002）也提出了很重要的一点，即高危群体是多基因座基因分型鉴定的，基于素质特异机制的干预可能只能解决他们风险度的特定一小部分。

确定遗传易感性对鉴定相关生活方式因素会有很大帮助。对已知全部是遗传易感的群体，这些生活方式因素更容易鉴定。这些信息会产生什么影响，意见分歧很大。两个极端见解可以画成积极的与消极的漫画（图 15.9）。积极的见解设想不但能够出现很强的保护性生活方式的改变，而且人们会听从建议而采纳它们。考虑到大多数发达国家目



图 15.9 积极的与消极的——关于鉴定复杂疾病易感因素对 21 世纪医学影响的观点对比
漫画来自 Maya Evans。

前的流行病，那些由吸烟、肥胖症和缺乏运动引起的所有可预防的疾病，这看起来是有点乐观的。而且，不应该完全陷入悲观：即使只有一种常见病能够被有效地预防，所付出的努力也是值得的。

（赵彦艳　译）

进一步阅读

Cardon LR, Bell JI (2001) Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99.

Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Longman, Harlow.

Kety SS, Rowland LP, Sidman RL, Matthysse SW (eds). (1983) *Genetics of Neurological and Psychiatric Disorders*. Raven Press, New York.

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd Edn. Johns Hopkins University Press, Baltimore.

Pollen DA (1996) *Hannah's Heirs* (expanded edition). Oxford University Press, Oxford.

Rosenthal D, Kety SS (1968) *The Transmission of Schizophrenia*. Pergamon Press, Oxford.

Sham S, Zhao J (1998) Linkage analysis using affected sib-pairs. In: *Guide to Human Genome Computing*, 2nd Edn (ed. MJ Bishop). Academic Press, San Diego CA.

Terwilliger J, Ott J (1994) *Handbook for Human Genetic Linkage*. Johns Hopkins University Press, Baltimore.

参考文献

ACMG/ASHG Working Group (1995) Use of ApoE testing for Alzheimer disease. *J.A.M.A.* **274**, 1627–1629.

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211.

Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**, 936–950.

Altshuler D, Hirschhorn JN, Klannemark M *et al.* (2000a) The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–79.

Altshuler D, Daly M, Kruglyak L (2000b) Guilt by association. *Nature Genet.* **26**, 135–137.

Amiel J, Lyonnet S (2001) Hirschsprung disease, associated syndromes and genetics: a review. *J. Med. Genet.* **38**, 729–739.

Arnos CI, Dawson DV, Elston RC (1990) The probabilistic determination of identity by descent sharing for pairs of relatives from pedigrees. *Am. J. Hum. Genet.* **47**, 842–853.

Badner JA, Sieber WK, Garver KL, Chakravarti A (1990) A genetic study of Hirschsprung disease. *Am. J. Hum. Genet.* **46**, 568–580.

Barsh GS, Farooqi IS, O'Rahilly S (2000) Genetics of body-weight regulation. *Nature* **404**, 644–651.

Byerley WF (1989) Genetic linkage revisited. *Nature* **340**, 340–341.

Camp NJ, Neuhausen SL, Tiobech J *et al.* (2001) Genomewide multipoint linkage analysis of seven extended Palauan pedigrees with schizophrenia, by a Markov-chain Monte Carlo method. *Am. J. Hum. Genet.* **69**, 1278–1289.

Carrasquillo MM, McCalion AS, Puffenberger EG, Kashuk CS, Nouri N, Chakravarti AS (2002) Genome-wide association study and mouse model identify interaction between *RET* and *EDNRB* pathways in Hirschsprung disease. *Nature Genet.* **32**, 237–244.

CHEK2-Breast Cancer Consortium (2002) Low-penetrance susceptibility to breast cancer due to *CHEK2**1100delC in non-carriers of *BRCA1* or *BRCA2* mutations. *Nature Genet.* **31**, 55–59.

Cox NJ (2001) Challenges in identifying genetic variation affecting susceptibility to type 2 diabetes: examples from studies of the calpain-10 gene. *Hum. Mol. Genet.* **10**, 2301–2305.

Deng H-W, Deng H, Liu Y-J *et al.* (2002) A genome-wide linkage scan for quantitative-trait loci for obesity phenotypes. *Am. J. Hum. Genet.* **70**, 1138–1151.

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322.

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001) Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nature Genet.* **28**, 361–364.

Emahazion T, Feuk L, Jobs M *et al.* (2001) SNP association studies in Alzheimer's disease highlight problems for complex disease linkage analysis. *Trends Genet.* **17**, 407–413.

European Consortium for IDDM Genome Studies (2001) A genomewide scan for Type 1-diabetes susceptibility in Scandinavian families: identification of new loci with evidence of interactions. *Am. J. Hum. Genet.* **69**, 1301–1313.

Feitosa MF, Borecki IB, Rich SS *et al.* (2002) Quantitative-trait loci influencing body-mass index reside on chromosomes 7 and 13: the National Heart, Lung and Blood Institute family study. *Am. J. Hum. Genet.* **70**, 72–82.

Fischer M, Harvald B, Hauge M (1969) A Danish twin study of schizophrenia. *Br. J. Psychiatr.* **115**, 981–990.

Fisher RA (1918) The correlation between relatives under the supposition of mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433.

Fodor FH, Weston A, Bleiweiss IJ *et al.* (1998) Frequency and carrier risk associated with common *BRCA1* and *BRCA2* mutations in Ashkenazi Jewish breast cancer patients. *Am. J. Hum. Genet.* **63**, 45–51.

Ford D, Easton DF, Stratton M *et al.* (1998). Genetic heterogeneity and penetrance analysis of the *BRCA1* and *BRCA2* genes in breast cancer families. *Am. J. Hum. Genet.* **62**, 676–689.

Fullerton SM, Bartoszewicz A, Ybazeta G *et al.* (2002). Geographic and haplotype structure of candidate Type 2 diabetes-susceptibility variants at the *Calpain-10* locus. *Am. J. Hum. Genet.* **70**, 1096–1106.

Gabriel SB, Salomon R, Pelet A *et al.* (2002a). Segregation at three loci explains familial and population risk in Hirschsprung disease. *Nature Genet.* **31**, 89–93.

Gabriel SB, Schaffner SF, Nguyen H *et al.* (2002b) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.

Göring HHH, Terwilliger JD, Blangero J (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.* **69**, 1357–1369.

Gulcher JR, Kong A, Stefansson K (2001) The role of linkage studies for common diseases. *Curr. Opin. Genet. Dev.* **11**, 264–267.

Gurling HMD, Kalsi G, Brynjolfson J *et al.* (2001) Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia on chromosomes 11q23.3-24 and 20q12.1-11.23. *Am. J. Hum. Genet.* **68**, 661–673.

Hampe J, Cuthbert A, Croucher PJP *et al.* (2001) Association between insertion mutation in *NOD2* gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925–1928.

Horikawa Y, Oda N, Cox NJ *et al.* (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26**, 163–175.

Hugot J-P, Chamaillard M, Zouali H *et al.* (2001) Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603.

IBD International Genetics Consortium (2001) International collaboration provides convincing linkage replication in complex disease through analysis of a large pooled data set: Crohn disease and chromosome 16. *Am. J. Hum. Genet.* **68**, 1165–1171.

Kendler KS, Gruenberg AM, Kinney DK (1994) Independent diagnoses of adoptees and relatives as defined by DSM-III in the provincial and national samples of the Danish Adoption Study of Schizophrenia. *Arch. Gen. Psychiatr.* **51**, 456–468.

Kety SS, Wender PH, Jacobsen B *et al.* (1994) Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen Study in the rest of Denmark. *Arch. Gen. Psychiatr.* **51**, 442–455.

Krawczak M, Schmidtke J (1994) *DNA Fingerprinting.* BIOS Scientific Publishers, Oxford, p. 650

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. *Am. J. Hum. Genet.* **57**, 439–454.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144.

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247.

Lander ES, Schork N (1994). Genetic dissection of complex traits. *Science* **265**, 2037–2048.

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**, 799–810.

Lewis CM, Levinson DF, Wise LH *et al.* (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, Part II: schizophrenia. *Am. J. Hum. Genet.* **73**, 34–48.

Luo D-F, Bui MM, Muir A, Maclaren NK, Thomson G, She J-X (1995) Affected sib-pair mapping of a novel susceptibility gene to insulin-dependent diabetes mellitus (*IDDM8*) on chromosome 6q25-q27. *Am. J. Hum. Genet.* **57**, 911–919.

Macdonald AS, Pritchard DJ (2001) Genetics, Alzheimer's disease and long-term care insurance. *North American Actuarial J.* **5**, 54–78.

McGuffin P (1984) In: *The Scientific Principles of Psychopathology.* (eds P McGuffin, MF Shanks, RJ Hodgson). Grune and Stratton, London.

McGuffin P, Huckle P (1990) Simulation of Mendelism revisited: the recessive gene for attending medical school. *Am. J. Hum. Genet.* **46**, 994–999.

Meyer MR, Tschanz JT, Norton MC *et al.* (1998) ApoE genotype predicts when – not whether – one is predisposed to develop Alzheimer disease. *Nature Genet.* **19**, 331–332.

Nathanson KL, Weber BL (2001) 'Other' breast cancer susceptibility genes: searching for more holy grail. *Hum. Mol. Genet.* **10**, 715–720.

Newman B, Austin MA, Lee M, King MC (1988) Inheritance of human breast cancer: evidence for autosomal dominant transmission in high-risk families. *Proc. Natl Acad. Sci. USA* **85**, 3044–3048.

Ogura Y, Bonen DK, Inohara N *et al.* (2001) A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606.

Onstad S, Skre I, Torgersen S, Kringlen E (1991) Twin concordance for DSM-III-R schizophrenia. *Acta Psychiatr. Scand.* **83**, 395–401.

Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. *Nature Rev. Genet.* **1**, 182–190.

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genet.* **31**, 33–36.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14.

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228.

Puffenberger E, Kauffman E, Bolk S *et al.* (1994a) Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on human chromosome 13q22. *Hum. Mol. Genet.* **3**, 1217–1225.

Puffenberger EG, Hosoda K, Washington SS *et al.* (1994b) A missense mutation of the endothelin-B receptor gene in multigenic Hirschsprung's disease. *Cell* **79**, 1257–1266.

Reich DE, Cargill M, Bolk S *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* **411**; 199–204.

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510.

Risch N (1990a) Linkage strategies for genetically complex traits. 1. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228.

Risch N (1990b) Linkage strategies for genetically complex traits. 2. The power of affected relative pairs. *Am. J. Hum. Genet.* **46**, 229–241.

Risch N (1990c) Linkage strategies for genetically complex traits. 3. The effect of marker polymorphism on analysis of affected relative pairs. *Am. J. Hum. Genet.* **46**, 242–253.

Risch N, Botstein D (1996) A manic depressive history. *Nature Genet.* **12**, 351–353.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 [see also *Science* **275**, 1327–1330 (1997) for discussion].

Risch N, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Res.* **8**, 1273–1288.

Schaid DJ (1998) Transmission disequilibrium, family controls and great expectations. *Am. J. Hum. Genet.* **63**, 935–941.

Sham PC, Curtis D (1995) An extended transmission disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* **59**, 323–336.

Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458.

Todd JA, Bell JI, McDevitt HO (1987) HLA-DQβ gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* **329**, 599–604.

Trimble MR (1996) *Biological Psychiatry*, 2nd Edn. John Wiley, Chichester, p. 183.

Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L (2000) Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur. J. Hum. Genet.* **8**, 604–612.

Weinberger DR, Egan MF, Bertolino A *et al.* (2001) Prefrontal neurons and the genetics of schizophrenia. *Biol. Psychiatry* **50**, 825–844.

Weiss KM, Terwilliger JD (2000) How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157.

Wilson JF, Goldstein DB (2001) Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am. J. Hum. Genet.* **67**, 926–935.

Wright AF, Carothers AD, Pirastu M (1999) Population choice in mapping genes for complex diseases. *Nature Genet.* **23**, 397–404.

Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H (2003) A polygenic basis for late-onset disease. *Trends Genet.* **19**, 97–106.

Yamazaki K, Takazoe M, Tanaka T, Kazumori T, Nakamura Y (2002) Absence of mutation in the *NOD2/CARD15* gene among 483 Japanese patients with Crohn's disease. *J. Hum. Genet.* **47**, 469–472.

第16章 分子病理学

本章内容

16.1 概述

分子病理学寻求对于一个既定的遗传改变为什么会导致某种特殊的临床表型的解释。我们在第11章里已经回顾了突变的性质和机制（概括于框16.1中）；本章主要关注它们对于表型的影响。分子病理学要求我们找出一个突变对于基因产物的数量或功能的影响，并且解释这个改变对特定的细胞、组织或者发育阶段为什么有或无致病性。

不足为奇的是，由于基因间相互作用的复杂性，分子病理学尚属一门相当不成熟的学科。迄今为止，最大的成就在于对肿瘤（其需要解释的表型——失控的细胞增殖——相对较为简单）以及血红蛋白病（其病理为珠蛋白异常的直接结果）的理解上。对于大多数的遗传病来讲，临床症状是一长串病因的最终结果。而分子病理学中所要追求的终极目标，即**基因型-表型对应**（genotype-phenotype correlation）将永远无法实现，因为事实上，即使"简单的"孟德尔疾病一点儿也不简单（Scriver and Waters，1999；Dipple and McCabe，2000；Weatherall，2001）。这些综述，尤其是由Scriver和Waters的论文是被极力推荐的（进一步阅读）。

框 16.1　突变的主要分类

范围从 1 bp 到数百万碱基对的**缺失**（deletion）

包括重复在内的**插入**（insertion）

单碱基替换（single base substitution）：

错义突变（missense mutation）在基因产物中将一个氨基酸替换为另一个

无义突变（nonsense mutation）将一个氨基酸的密码子替换为一个终止密码子

剪接位点突变（splice site mutation）产生或破坏外显子-内含子剪接的信号

移码（frameshift）可由缺失、插入或剪接错误产生

动态突变（dynamic mutation）为在传递给子女时常改变长度的串联重复

见表 16.1 中的一些例子，以及第 11 章中的更多细节和关于机制的讨论。

尽管存在多种困难，对于分子病理学的学生来讲，人类有一个巨大的优势：我们对于人类表型的了解比对其他任何生物都多得多。这不仅使我们更容易注意到人类而非果蝇或蠕虫的微小变异，而且全球范围的医疗体系构成了一个庞大且持续的突变筛查体系。任何发生频率大于 $1/10^9$ 的人类表型都可能已经在文献中被描述过。此外，对于大部分已经鉴定的疾病基因，许多不同的突变已被发现，我们不可能用人类做实验，或有目的地繁育他们。但人类为观察一个特定基因的许多不同的改变对于表型的影响提供了独特的机会。当人类基因组计划的重点从编制基因目录转移到了解它们的功能时，分子病理学的研究就走到舞台的中央。对于人的研究将产生假说，而后者必须在动物上被验证。因此，对自然发生的人类突变的研究将从对于动物中特异性的自然或人工突变的研究中得到补充。

16.2　*A* 和 *a* 等位基因的简便命名中暗藏了巨大的 DNA 序列多样性

当我们将一名囊性纤维化携带者的基因型描述为 *Aa* 时，我们用 *a* 来表示任何 *CFTR* 基因序列的突变，使它因此不能产生一个功能性氯通道。已报道了超过 750 种不同的这类等位基因。同理，*A* 表示任何功能性序列——在非近亲的人群中，*A* 基因的实际 DNA 序列并不一定 100％相同。对于遗传咨询和系谱分析来说，这是最实用水平的描述，然而对分子病理学而言，我们需要更为仔细地观察。框 16.1 简要总结了对致病性突变中可见的 DNA 序列改变的主要类型。框 16.2 总结了描述它们的通常习惯。

分子病理学中存在的一个问题是，现在没有列举所有已知人类突变的完整数据库。人类突变数据库（http://www.hgmd.org）包含对一部分基因有用的列表，但并不完整，而 OMIM 在列举突变等位基因上也不详尽或一致。通过 PubMed 无法获得这类信息，因为科研杂志通常不登载已经研究得很深入的基因的新突变的报道。目前正在尝试通过 HUGO 突变数据库计划来解决这一问题（详情和许多有用的链接参见http://www.genomic.unimelb.edu.au/mdi/）。

框 16.2　描述序列变异的术语

详见http://www.dmd.nl/mutnomen.html及 den Dunnen 和 Antonarakis（2001）。

氨基酸替换

如有必要，以“p.”开头来表示蛋白质，采用单字母代码：

A. 丙氨酸	C. 半胱氨酸	D. 天冬氨酸	E. 谷氨酸	F. 苯丙氨酸	G. 甘氨酸
H. 组氨酸	I. 异亮氨酸	K. 赖氨酸	M. 蛋氨酸	N. 天冬酰胺	P. 脯氨酸
Q. 谷氨酰胺	R. 精氨酸	S. 丝氨酸	T. 苏氨酸	V. 缬氨酸	W. 色氨酸
Y. 酪氨酸	X 为终止密码子，也可以采用三字母代码				

p. R117H 或 Arg117His—117 位精氨酸被组氨酸替代（起始的蛋氨酸为密码子 1）

p. G542X 或 Gly542Stop—542 位甘氨酸被终止密码子替代

核苷酸替代（nucleotide substitution）

如有必要，以“g.”（基因组）或“c.”（cDNA）开头，起始密码子 ATG 的 A 为+1，它的前一个碱基为−1，没有 0，在核苷酸变异之前给出其顺序号。对于内含子中的变异，如果仅知道完整的 cDNA 序列，用内含子序号 1VSn 或距离最近的外显子的位置序号来注明。

g. 1162G→A—1162 位点鸟嘌呤被腺嘌呤所替代。

g. 621+1G→T 或 IVS4+1G→T—第 4 内含子第 1 个碱基由 G 变为 T；外显子 4 在 nt621 处结束。

缺失和插入（deletion and insertion）

用 del 表示缺失，ins 表示插入。同上，对于 DNA 变异，首先指出核苷酸位点或间隔，对于氨基酸变异，首先给出氨基酸的代码。

p. F508del——508 位苯丙氨酸缺失

c. 6232_6236del 或 c. 6232_6236delATAAG——由 cDNA 的 nt6232 处开始，缺失 5 个核苷酸（注明）。

g. 409-410insC——基因组 DNA 在 nt409 和 410 之间插入 C。

16.3　突变的一级分类：功能丢失性突变和功能获得性突变

16.3.1　对于分子病理学而言，重要的是突变等位基因的影响而非它的序列

了解突变等位基因的序列对于基因检测来讲十分重要（18 章），但对于分子病理学而言，我们需要知道它是干什么的。一个突变的基因对生物体可能有各种微妙的影响，但一个首要问题就是它是否造成功能的丢失或获得。

- 在**功能丢失性突变**（loss of function mutation）中，基因产物减少或无功能。
- 在**功能获得性突变**（gain of function mutation）中，基因产物具有某种明显的异常功能。

框 16.3 给出了一种描述这些效应的方法。

框 16.3　描述等位基因效应的术语

无效等位基因（null allele or amorph）：无产物的等位基因

亚效等位基因（hypomorph）：产物数量或活性降低的等位基因

超效等位基因（hypermorph）：产物数量或活性增加的等位基因

新效等位基因（neomorph）：产生新功能或新产物的等位基因

负效等位基因（antimorph）：其功能或产物可拮抗正常产物功能的等位基因

功能丢失性突变常常产生隐性表型。对于多数基因产物而言，精确的产量并非关键，正常数量的一半就够用了。因此，多数先天性代谢缺陷为隐性遗传。然而，对一些基因产物来说，正常水平的 50%无法满足正常功能的需要，**单倍性不足**（haploinsufficiency）将产生某种异常表型，因此以显性方式遗传（节 16.4.2）。有时，杂合子个体中一个无功能性突变多肽将影响正常等位基因的功能，产生一种**显性负效应**（dominant negative effect）（框 16.3 内术语中的反效等位基因，见节 16.4.3）。

因为一个正常等位基因不能阻止突变等位基因异常工作，功能获得性突变通常导致显性表型。这常常涉及某种控制或信号系统的功能异常——在适当的时候发出信号，或在应该关闭某个进程时却没有关闭，有时候功能获得性涉及其产物发挥了新的功能——例如，某种含有扩增的谷氨酰胺重复的蛋白将发生异常聚集。

不可避免的是，一些突变并不能被简单地划分为功能丢失或功能获得。一个永久开放的离子通道是失去了关闭的功能，还是获得了不恰当的开启功能？显性负性突变等位基因表示它的功能丢失了，但也具有一些明显异常的功能。一个突变可能破坏基因产物的几种功能间的平衡。尽管如此，鉴别功能丢失和功能获得是思考分子病理学的首要工具。

16.3.2　当一个基因的点突变产生与缺失一样的病理变化时，可能是功能丢失

在缺乏生化研究的情况下，纯粹的遗传学证据常常可以指示一种表型是由功能丢失或获得所致。当某种临床表型由一个基因的功能丢失所致时，我们会发现灭活基因产物的任何改变都会产生相同的临床表型。我们将可能找到与缺失或破坏该基因的突变效应相同的点突变。Waardenburg 综合征 1 型（MIM 193500：听力丧失及色素异常）就是一个例子。

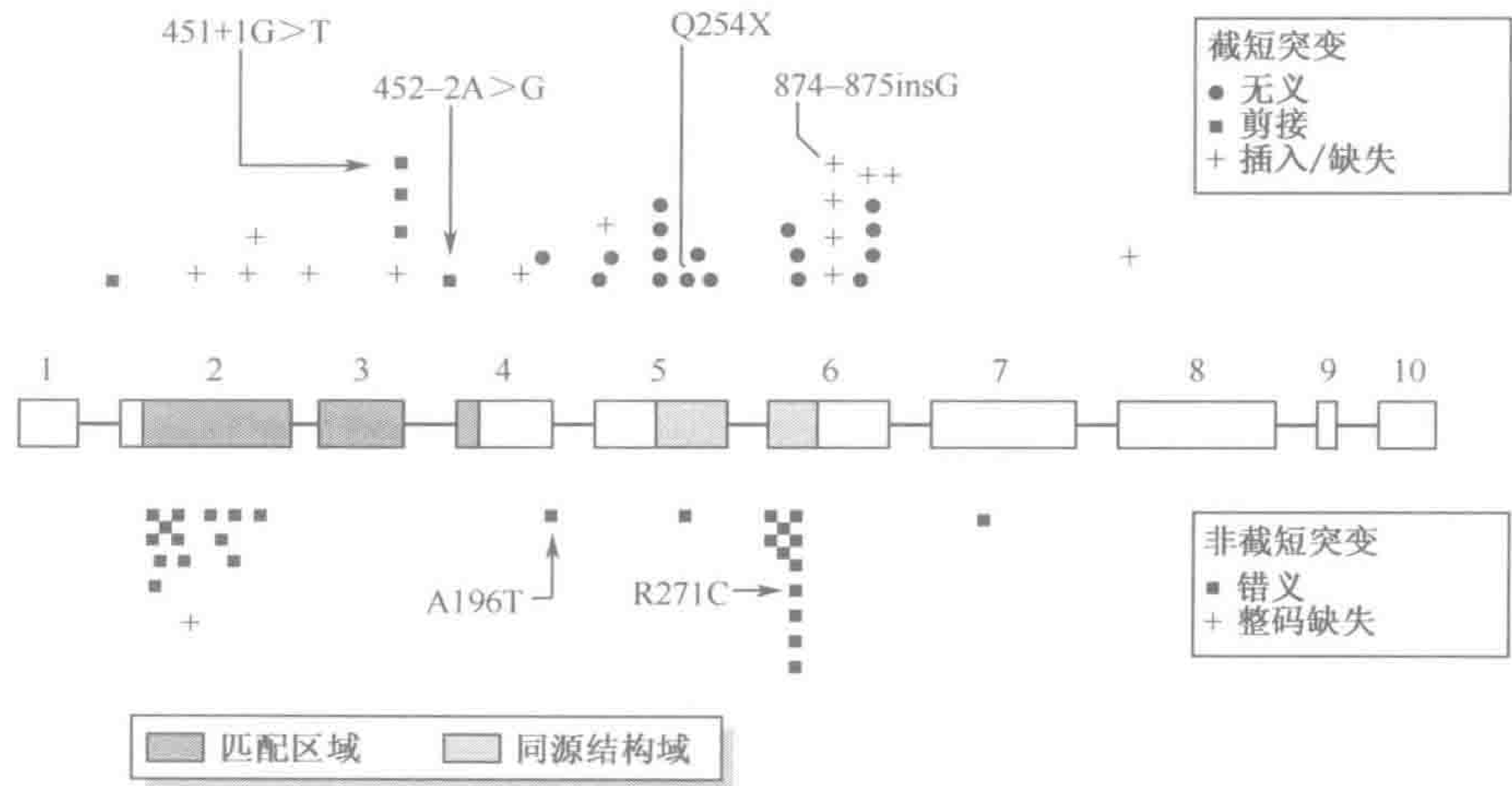


图 16.1　*PAX3* 基因的功能丢失性突变

基因的 10 个外显子以框来表示，接连的内含子未按比例表示，阴影区域表示 PAX3 蛋白的两个 DNA 结合域的编码序列，值得注意的是，完全破坏 PAX3 蛋白结构的突变（画在基因图上方）分散在基因的至少前 6 个外显子中，而错义突变（画在基因图下方）集中在两个区域，匹配结构域的 5′部分及同源结构域（homeodomain）的第三个螺旋。A196T 据信可以影响剪接，其他已命名的突变将在表 16.1 中被提示。874-875insG 突变在一串 6 个 G 中引入第 7 个 G。它已经独立地发生了多次，提示了相对频繁的滑链错配（节 11.3.1）。

如图 16.1 所示，*PAX3* 基因的致病突变包括氨基酸的替代、移码、剪接突变及在一些患者中的基因的完全缺失。由于所有这些都产生同样的临床表现，其病因必然是 *PAX3* 基因的功能丢失。同样，在由不稳定三核苷酸重复所致的疾病中（节 16.6.4），脆性 X 染色体和 Friedreich 共济失调有时是由于它们各自基因的其他类型的突变所致，指示是功能丢失，而 Huntington 病从未见任何其他类型的突变，指示是功能获得。

16.3.3　功能获得可能仅发生于一个基因的某种特殊突变才能产生的特定病理

与功能丢失相比，功能获得可能需要更特殊的变异。功能获得状态下的突变谱应该相应地更为有限，而基因的缺失或破坏应无法产生同样的情况。可能的例子包括 Huntington 病（节 16.6.4）以及软骨发育不全（MIM 100800：短肢侏儒）。几乎所有的软骨发育不全都具有同样的氨基酸改变，即成纤维细胞生长因子受体 FGFRS 中的 G380R，该蛋白质的其他替代将产生其他综合征（节 16.7.3）。由于原因不明，尽管需要一个非常特殊的 DNA 序列变异，G380R 的突变率异常地高，因而，软骨发育不全是较常见的遗传病之一。

突变同质性（mutational homogeneity）是功能获得的一个指标，然而也有其他原因能够解释为什么单一突变是某种疾病大多数或全部病例的病因：

- 所观察的疾病与基因产物本身直接相关。而不是遗传变异的更为遥远的结果，这类疾病可以被界定为提示变异的产物。如镰形细胞病（见框 16.4）；
- 某些特殊的分子机制使得在基因中可能更易于发生特定的序列改变，例如脆性 X 综合征中的 CGG 扩增（节 16.6.4）；
- 可能存在某种**建立者效应**（founder effect）——例如，在 Ashkenazi 犹太人中特定疾病的突变很常见，它可能反映了存在于现代 Ashkenazi 人群的一个非常少数的建立者中的突变（Motulsky，1995）；
- 有利于杂合子的选择（节 4.5.3）增强了建立者效应，并且常常造成一种或少数特殊突变在一个人群中变得常见。

16.3.4　判断某一 DNA 序列改变是否为致病性较为困难

在受累者中发现的每个序列变异并非一定是致病性的，我们如何判断所发现的序列变异正是要找的致病性突变或是无害的变异呢？按递降可信度排列，标准如下：

1）功能研究表明该变异为致病性；

2）先例：以往曾在患该病的患者中发现过该变异（而在匹配的同种族对照中不存在）；

3）新突变，在双亲中不存在，但在具有新发疾病者中存在；

4）新的序列变异，在 1 队，如 100 个正常对照中不存在；

5）序列的性质改变（框 16.5）。

功能研究是金标准，其他任何标准都并非完全可信。如上所述，对于功能获得性表型，其原因可能非常特殊。任何与标准突变不同的序列变异都可能并非为致病性，至少对于研究的疾病而言，功能丢失的情况在解读上则提出了多得多的问题。

16.4　功能丢失性突变

16.4.1　基因的许多不同改变均能导致功能丢失

毫不奇怪的是，有许多方式可以降低或破坏基因产物的功能（表 16.1）。其中的一些途径已在节 11.4 中讨论过。血红蛋白病（框 16.4）很好地例证了多种这类机制。事实上，珠蛋白可以用于示范本书中描述的几乎各种过程，推荐读者查阅如 Weatherall 等（2001；见进一步阅读）所著的综述中相应的分子病理学叙述。

框 16.5 给出了判断一种突变对于基因产物的可能作用的一些原则。

小的缺失及插入如果造成移码（即如果增加或减少并非 3 的整数倍的核苷酸），那将对基因产物产生显著的影响。杜兴肌营养不良基因的缺失是个明显的例子（图 16.2）。通常，移码缺失将导致严重的杜兴肌营养不良，其中无抗肌萎缩蛋白产生，而非移码突变则将导致较轻微的 Becker 型，其中存在抗肌萎缩蛋白但不正常。

表 16.1　减少或消除功能性基因产物产生的 11 种途径

改变	例子
缺失：	
（Ⅰ）整个基因	大部分 α 地中海贫血突变（图 16.3）
（Ⅱ）基因的部分	60% Duchenne 肌营养不良（图 16.2）
基因中插入一段序列	甲型血友病 *F8* 基因中插入 LINE-1 重复序列（节 11.5.6）
破坏基因的结构	
（Ⅰ）易位	Duchenne 肌营养不良妇女中的 X-常染色体易位（图 14.10）
（Ⅱ）倒位	*F8* 基因中的倒位（图 11.20）
阻止启动子工作	
（Ⅰ）突变	β 珠蛋白 g.-29A→G 突变（表 18.5）
（Ⅱ）甲基化	许多肿瘤中的 *CDKN2A* 基因（节 17.6.1）
使 mRNA 不稳定	
（Ⅰ）多聚腺苷酸位点突变	α 珠蛋白 g. AATAAA →AATAGA 突变
（Ⅱ）无义突变介导的 RNA 降解	β 珠蛋白 p. Q39X
阻止正常剪接（节 11.4.3）	
（Ⅰ）失活供体剪接位点	*PAX3* g.451＋1G→T 突变（图 16.1）
（Ⅱ）失活受体剪接位点	*PAX3* g.452－2A→G 突变（图 16.1）
（Ⅲ）改变某个外显子的剪接增强子	*SMN2* 外显子 7 g.C6T（Cartegni and Krainer，2002）
（Ⅳ）激活某个潜在的剪接位点（可能位于某个内含子的深处）	*LGMD2A* G624G（图 11.12）
	β 珠蛋白 IVS1－110G→A 突变（表 18.5）
	*CFTR*3849＋10kb C→T（表 18.6）
在翻译过程中引入一个移码	*PAX3* g.874-875insG 突变（图 16.1）
将某个密码子变为终止密码子	*PAX3* p.Q254X 突变（图 16.1）
替换一个重要的氨基酸	*PAX3* p.R271C 突变（图 16.1）
阻止转录后的加工	Ehlers Danlos Ⅶ综合征中抗分裂胶原 N 端前肽（节 16.6.1）
阻止产物在细胞内正确定位	囊性纤维化中 p.F508del 突变

框 16.4 血红蛋白病

由于许多原因，血红蛋白病在临床遗传学中占有特殊的地位，它们是迄今世界范围内最常见的严重性孟德尔疾病。珠蛋白展现了基因组进化以及人群中疾病的种种重要方面（图 12.4～12.6）。关于血红蛋白的突变和疾病的描述比其他基因家族多得多（图 10.22，10.23）。临床症状与蛋白质的功能异常直接相关，后者在每 100ml 血液中为 15g，很容易被研究，因此，分子和临床变化间的关系对于血红蛋白病来讲比其他大部分疾病更清晰。

血红蛋白病分为三个主要类型

- 地中海贫血（thalassemias）是由 α 和 β 链数量不足所致。等位基因可以被分为无产物（α^0，β^0）和产生数量降低的产物（α^+，β^+）。致病的缺陷包括在节 16.4 中详细描述的全部例子（见表 18.5）
- 异常血红蛋白（abnormal hemoglobin）是由于氨基酸变异导致的多种疾病，其中镰形细胞贫血最著名。E6V 突变是 β 珠蛋白外侧表面的极性氨基酸被中性氨基酸所替代，这增加了分子间的黏附，导致脱氧血红蛋白的聚集及红细胞的破坏，镰形红细胞的存活时间下降（导致贫血），并易于阻塞毛细血管，使梗塞处下游的器官发生缺血和梗塞。其他的氨基酸改变能够导致贫血、紫绀、红细胞增多症（红细胞数量过多）和高铁血红蛋白血症（铁从亚铁转换为三价铁态）等。
- 遗传性胎儿血红蛋白持续增高症（hereditary persistence of fetal hemoglobin）是由于从胎儿到成人血红蛋白的正常转换的缺陷所致，是其他两类突变体临床效应的重要修饰因素。

见 Weatherall 等（2001；进一步阅读）的详细综述。

框 16.5 评估 DNA 序列变异重要性的原则

- 整个基因的缺失，无义突变以及移码将几乎肯定破坏基因的功能。
- 大部分内含子中旁侧序列保守的 GT…AG 核苷酸的突变将影响剪接，并将消除基因的功能。许多其他的序列改变也能够影响剪接，但其途径则要难以预测得多（见下文）。
- 错义突变如果影响了一个蛋白中的功能重要部分，则更可能是致病性的，对于蛋白质结构的计算机建模，有助于提示哪些残基更为重要。例如，图 16.1 所示的所有无义突变可导致功能丢失，均集中在 PAX3 蛋白质关键的 DNA 结合区域。
- 如果某个氨基酸在有亲缘关系的基因中保守，该氨基酸的改变则可能影响功能（种间同源或种内同源）。
- 氨基酸替换如果为非保守性（将一个极性氨基酸替换为非极性的，或一个酸性氨基酸替换为碱性的；框 11.3），将更可能影响功能。

无义突变［即产生**提前终止密码子**（premature termination codon）的突变］的功能通常等同于无效等位基因，因为它们将引起**无义突变介导的 mRNA 降解**（nonsense-mediated mRNA decay）（Hentze and Kulozik，1999）。mRNA 很少被翻译而产生一种截短的蛋白质。当提前的终止密码子位于最后一个剪接连接处上游至少 50 个核苷酸时，通常会出现无义突变介导的降解。这一机制的细节见 Lykke-Andersen 等（2001）。

剪接突变通常被认为就是改变内含子末端保守的 GT…AG 序列的突变。事实上种类丰富得多的序列变化均可影响剪接，但它们的影响都难以预测。因此，许多影响剪接的突变在用 RT-PCR 分析之前均未被识别。因此，对于这类突变存在有诊断不足。在

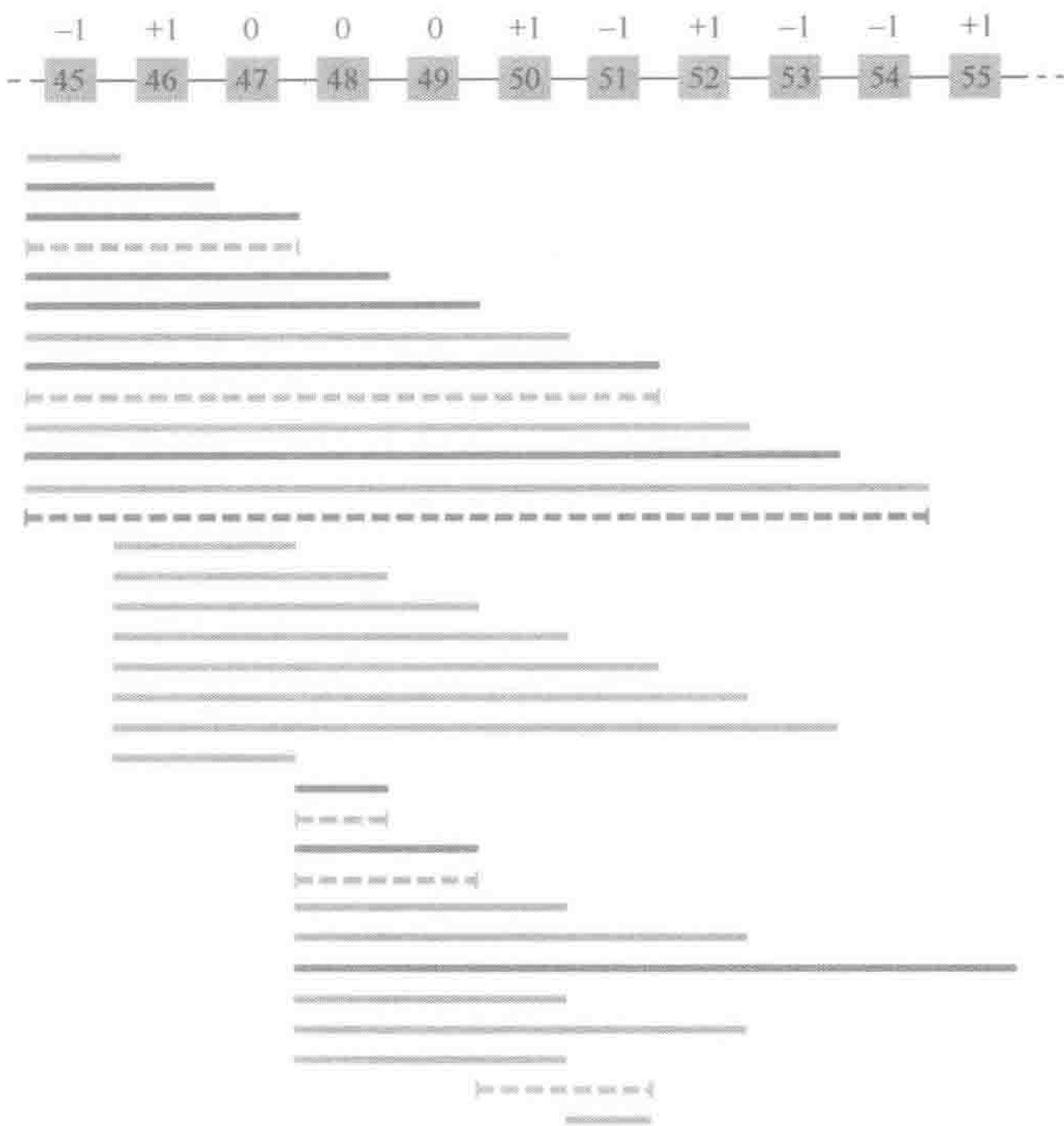


图 16.2　Duchenne 或 Becker 肌营养不良患者中抗肌萎缩蛋白基因中央部分的缺失

编号的方块代表外显子 45～55，方块上方的数字表示缺失该外显子对读框产生的影响（0 无效，-1 示向后移动一个核苷酸，+1 示向前移动一个核苷酸）。灰线条表示致死性 DMD 患者所携带的缺失，黑线条代表较轻微的 BMD 患者所携带的缺失。通常移码缺失将导致 DMD，读框中性缺失将导致 BMD。例外由虚线表示。例外的原因包括缺失终止于某个外显子中、修饰基因的作用或环境的影响，也可能是临床或实验诊断的错误。数据源自 Leiden 肌营养不良网址 http://www.dmd.nl，从中可以查询到更全面的数据及背景资料。图 18.6 示意在实验室中这些缺失是如何被鉴定的。

表 18.5 所列举的 β 珠蛋白突变中有若干例子。

三种类型的序列改变可以导致剪接异常：

- **正常剪接位点改变**：剪接（节 11.4.3）不仅需要标准的 GT…AG 序列，也需要该位点周围严格性稍低的序列，此处的变异可能改变不同剪接异构体的比例，而不是破坏所有的剪接——例如，*CFTR* 基因（见 MIM 602421）内含子 8 中的 5T/7T/9T 多态将改变跨越外显子 9 的转录体比例，这类影响可能是遗传疾病常见的原因，特别可能是在对常见疾病的易感性上（Nissim-Rafini and Kerem，2002）。
- **剪接增强子或沉默子的改变**：这些重要但未很好定义的序列可能位于外显子或内含子中（Nissim-Rafinia and Kerem，2002）。Cartegni 和 Krainer（2002）列举了一个特别明确地破坏了外显子的剪接增强子的突变的例子。
- **潜在剪接位点的激活**：看似无害的碱基替换可能使既往失活的序列被用作一个剪接位点。潜在的位点可能位于外显子或内含子中。图 11.12 举了一个例子。如果该位点深藏在内含子中，若不用 RT-PCR，该突变将可能被遗漏——例如，*CFTR* 突变 3849＋10kbC→T 激活了位于 19 内含子中 10kb 处的一个潜在剪接位点。

16.4.2 在单倍性不足中，基因功能水平下降 50 %将导致异常表型

由于杂合子常常完全正常工作，功能丢失性突变倾向于表现为隐性。有些时候这是由于转录或蛋白质活性水平的反馈环路对剂量减少的代偿。但在许多情况下，细胞和有机体在仅 50%水平的基因活性情况下也能正常行使功能，仅有相对极少的基因呈现**单倍性不足**；表 16.2 举出了一些实例。少数其他基因显现其他类型的**剂量敏感性**（dosage sensitivity）（图 16.3 和图 16.8）。

表 16.2 可能由单倍性不足所致的表型（详见正文）

疾病	MIM	基因	注释
Alagille 综合征	118450	*JAG1*	节 16.8.1
多发性外生骨疣	133700	*EXT1*	表 16.8
香肠样神经病变	162500	*PMP22*	节 16.6.2
瓣膜上主动脉狭窄	185500	*ELN*	本节
发-鼻-咽综合征	190350	*TRPS1*	表 16.8
Waardenburg 综合征 1 型	193500	*PAX3*	节 16.3.2

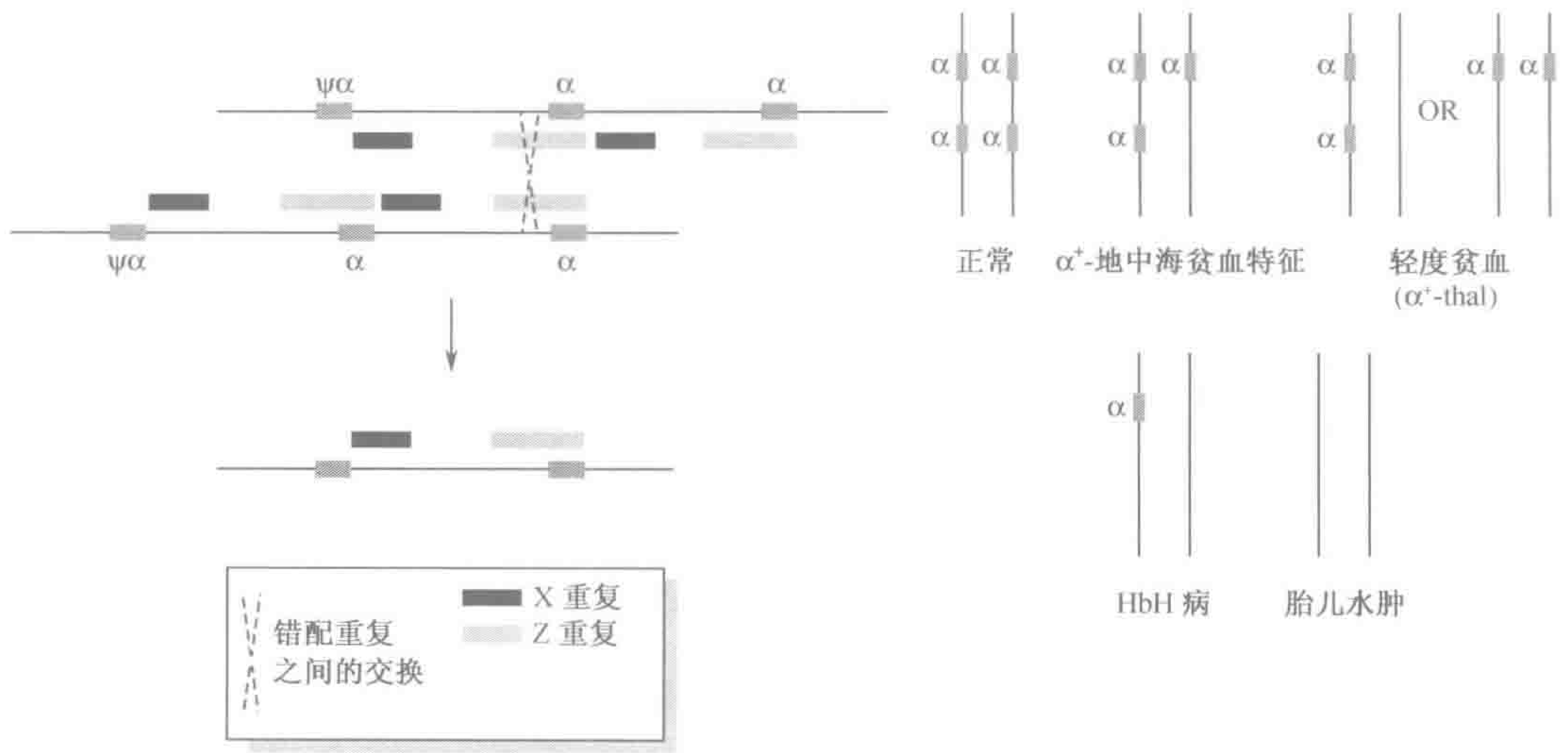


图 16.3 α 地中海贫血中 α 珠蛋白基因的缺失

正常的 16 号染色体携带串联排列的两个活性 α 珠蛋白基因和 1 个失活的假基因，重复区（标记为 X 和 Z）可以错排，容许不等交换。该图显示，错排的重复的不等交换是一条仅具有一个活性 α 基因的染色体。X 重复之间的不等交换具有相似的效应，其他重复（未显示）之间的不等交换可产生携带无功能性 α 基因的染色体。个体因此可能具有 0～4 个或者更多的 α 珠蛋白基因。其后果随 α 基因数量减少而逐渐严重。详见 Weatherall 等（进一步阅读）。

人们可能有理由会问，为什么任何基因产物都会有剂量敏感性。为什么自然选择没有使情况变得更好？如果一个基因的两个拷贝的表达产生刚刚足够数量的产物，那对于表达较高水平的变异体的选择，就应该演化出更加强壮的生物体而不必花费明显的代

价，答案是，大多数情况下这的确发生过。这就是为什么相对较少的基因为剂量敏感性。然而，特定的基因功能本来就有剂量敏感性。它们包括：

- 基因产物为数量信号传导系统的一部分，其功能依赖于对受体、DNA 结合位点等的部分或不同程度的占有；
- 基因产物互相竞争而决定某一发育或代谢开关；
- 基因产物以固定的化学计量方式在相互作用中互相配合（诸如 α 和 β 珠蛋白以及许多结构蛋白）。

在各种情况下，基因产物被细胞中的其他某种物质所滴定，重要的不是产物的正确的绝对水平，而是相互作用的产物的正确的相对水平。对于所有的相互作用的配体的改变，这种作用均很敏感，因此这些显性的状况常呈现高度可变的表现（节 16.6.3），产物基本上独立工作的基因，如代谢中的许多可溶性酶，极少表现出剂量效应。另外，在某些情况下，细胞只有一个正常工作的基因拷贝，发挥的功能不能满足对基因产物的大量需求，弹性蛋白就是一个例子。对于具有弹性蛋白基因的缺失或功能丢失性突变的杂合个体而言，仅需要中等数量弹性蛋白，组织（皮肤、肺脏）不受累，但主动脉的弹性蛋白多得多，将常常表现为需要手术治疗的异常（瓣膜上主动脉狭窄）。

16.4.3　以二聚体或多聚体方式工作的蛋白质中的突变有时会产生显性负效应

显性负效应（dominant negative effect）是指在一个杂合子中突变的多肽不仅失去其自身的功能，同时还对正常等位基因的产物造成干扰。有些人可能会认为这属于功能获得，而不是功能丢失性突变，但这仅是文字上的争论。显性负突变将造成比同一基因的单纯无效突变更为严重的后果，构成多聚体结构的结构蛋白尤其易受显性负效应的影响。胶原蛋白就是一个经典的例子。

胶原纤维是结缔组织中的主要结构蛋白，是由三条多肽链螺旋而成，有时为同源三体，有时为异源三体，它们组装成紧密包装的交互联结的阵列，形成强韧的纤维。在新合成的多肽链（原胶原）中，N 和 C 端的前肽的旁侧为规则重复序列 $(\text{Gly-X-Y})_n$，其中 X 或 Y 通常为脯氨酸，而另一个则是其他氨基酸。三条胶原链在 C 端前肽的控制下联合并缠绕成一个三螺旋。在三螺旋形成之后，N 和 C 端前肽被切割掉。与一条正常链复合，但随后破坏三螺旋结构的前肽，可以使功能性胶原的产量大大低于 50%（图 16.4）。胶原突变的分子病理学将在后面讨论（节 16.6.1 和表 16.7）。

二聚或寡聚非结构性蛋白质亦可以表现出显性负效应。例如，b-HLH-Zip 的转录因子家族（图 10.9）与 DNA 结合成为二聚体。无法形成二聚体的突变常导致隐性表型，但可将有功能的分子变为无活性的二聚体的突变，则表现为显性表型（Hemesath *et al.*,1994）。细胞膜的离子通道是易于发生显性负效应的多聚体结构的另一个例子（节 16.6.1）。

16.4.4　在无 DNA 序列改变的情况下，表观遗传修饰亦能使基因功能消失

不依赖于 DNA 序列改变的可遗传的变异（从细胞到子细胞、或者从父母到子女）被称为**表观遗传**（epigenetic）（节 10.1）。1998 年 5 月 1 日，*Cell* 上有一系列的短篇综

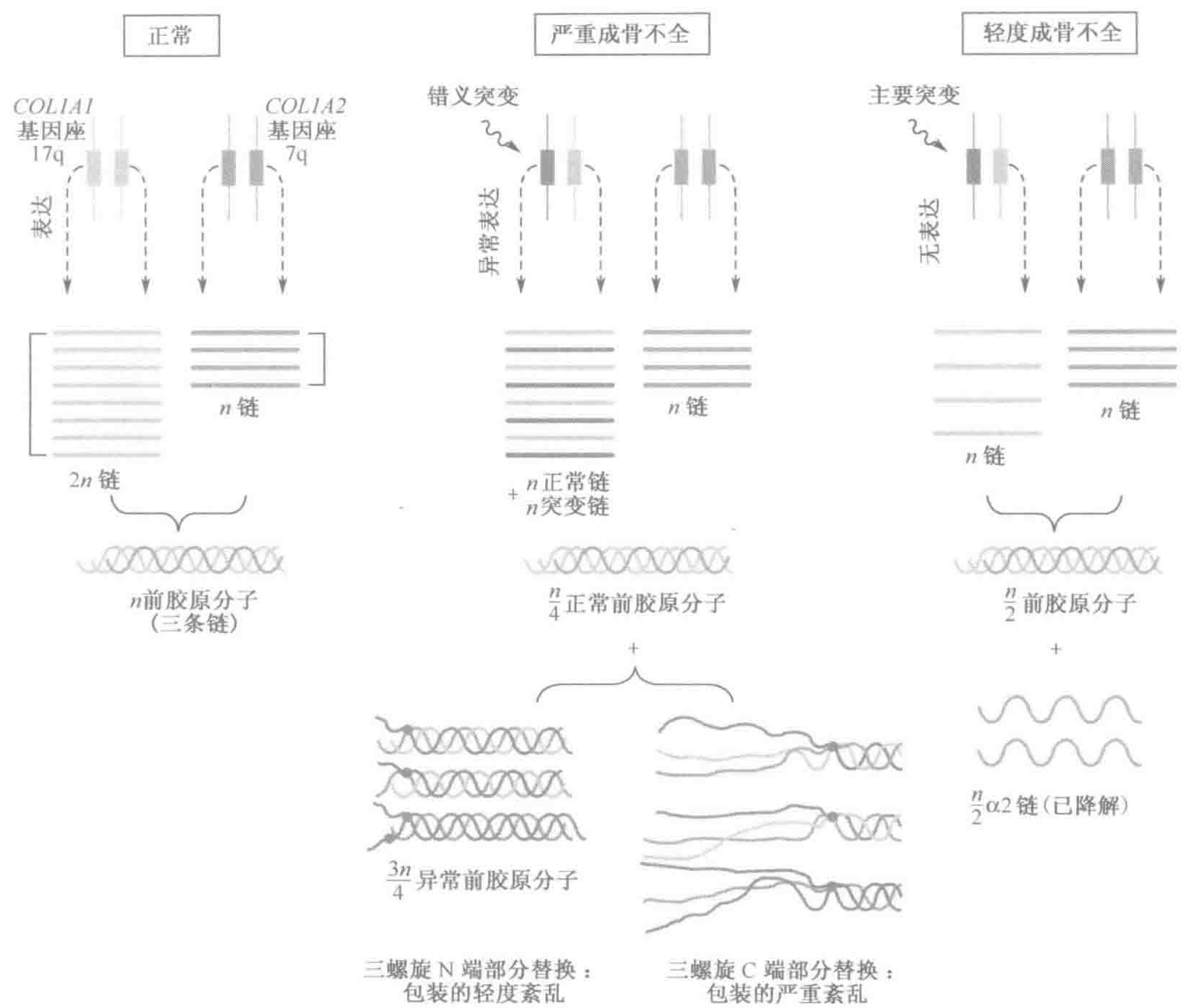


图 16.4　胶原基因突变的显性负效应

胶原细丝由三螺旋前胶原单元排列而成。Ⅰ型原胶原由 *COL1A1* 以及 *COL1A2* 编码，含有两条链。两者之一的无效突变比编码多肽链的突变严重性较轻，编码的多肽链整合进三螺旋并破坏其功能。

述讨论了表观遗传学的丰富内容的许多方面。癌症中的表观遗传效应尤为明显（Jones and Baylin，2002）。在这里，主要的表观遗传机制为 **DNA 甲基化**（DNA methylation），有关此话题的详细内容见 Bird（2002）的综述。

如前所述（节 10.4.2），人类 DNA 上与鸟嘌呤相邻的胞嘧啶碱基常常在 C_pG 双核苷酸的 5 位碳原子上被甲基化，而 C_pG 甲基化的模式在 DNA 复制时可稳定地传递（图 16.5），这些甲基化模式是控制基因转录（节 10.4.3）的重要信号。X 染色体的失活至少部分地依赖于 DNA 甲基化。而甲基化模式的表观遗传学传递则保证了同样的 X 染色体在母亲和女儿细胞中的失活。不适当的甲基化可导致遗传性的致病性功能丢失。例如，在许多肿瘤中，p16 抑癌基因（*CDKN2A*）的功能丢失是由于启动子甲基化而不是由于 DNA 序列的突变（节 17.4.3）。反过来，不适当的去甲基化被认为在肿瘤中启动了癌基因的表达。

印迹（imprint）基因（节 4.3.4 和 10.5.4）是表观遗传修饰的一个尤其有趣的例

子。它们的表达受控于甲基化模式，后者因基因亲代的来源而异。当印迹机制失效或亲代起源与预期的不同时，整个基因将发生功能丢失或不当表达。印迹基因成簇出现，既包括父源也包括母源印迹基因。它们中的一些仅在特定组织中印迹。它通常 DNA 的双链都可能被转录，一个转录物成为 mRNA，另一个大的（＞100kb）为非翻译的**反义 RNA**（antisense RNA），但一条链的转录将阻碍另一条的转录。因此，人类印迹疾病的分子病理学非常有趣却特别复杂，框 16.6 描述了最有名的人类印迹疾病。Prader-Willi 和 Angelman 综合征；有关其他疾病的详细信息见 OMIM 130650（Beckwith-Wiedemann 综合征），OMIM 139320（*GNAS*）及 Reik 和 Walter（2001）以及 Rougeulle 和 Heard（2002）的综述。

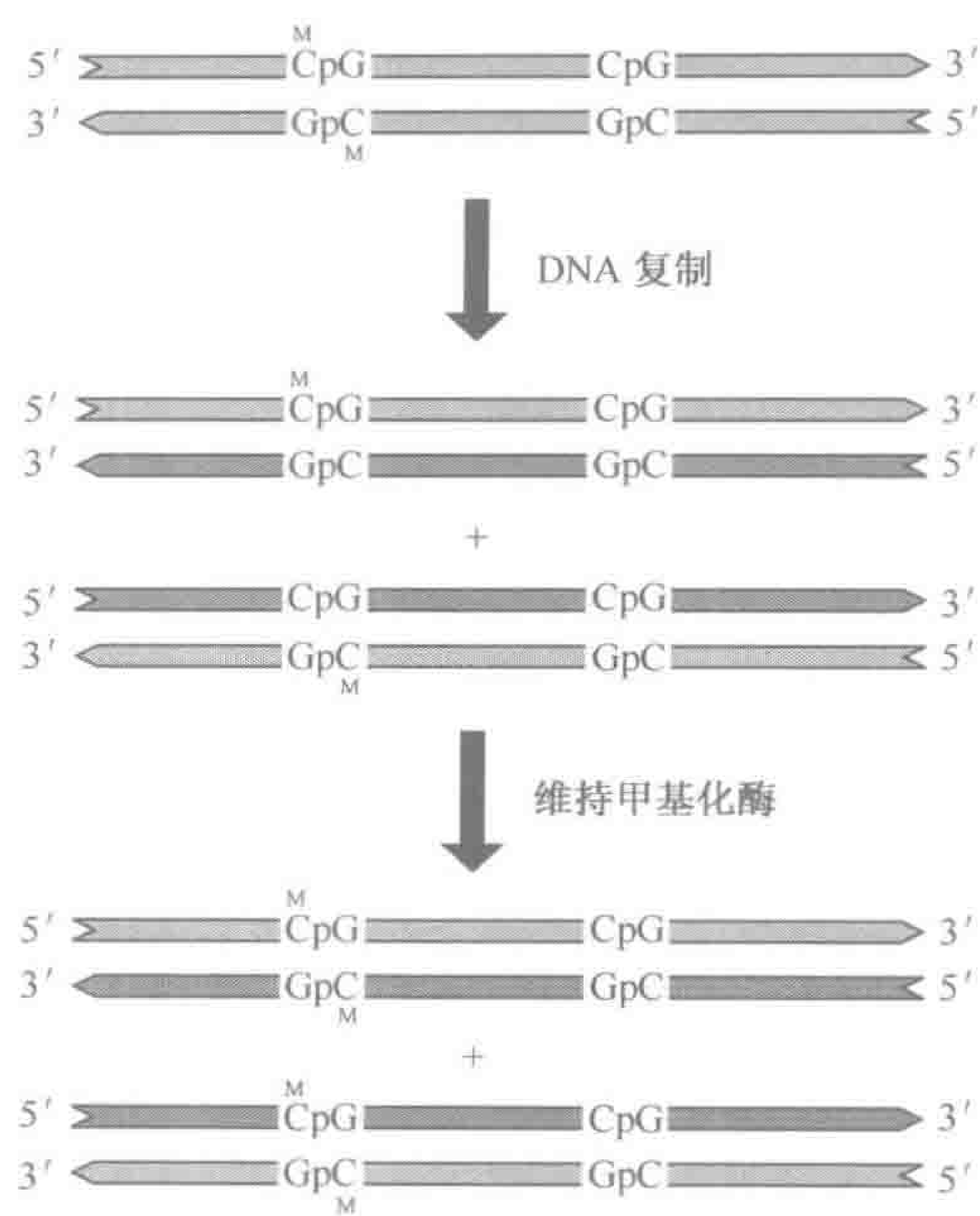


图 16.5　CpG 甲基化模式的遗传

维持甲基化酶可识别新复制 DNA 中的半甲基化 CpG，容许 CpG 的甲基化模式遗传下去［节 10.4.2 及 Bird（2002）］。

框 16.6　Prader-Willi 和 Angelman 综合征的分子病理学

Prader-Willi 综合征（PWS）和 Angelman 综合征（AS）均为 15q11-q13 上差异性印迹基因的异常所致。它们例示了印迹基因簇相关的复杂分子病理学（Nicholls and Knepper，2001）。

- PWS（MIM 176260：智力低下、低张力、整体性肥胖、男性性腺发育不全）是由仅自父源染色体表达的基因的功能缺失所致。
- AS（MIM 105830：智力低下、失语、生长迟滞、多动、失态大笑）是由于仅自母源染色体表达的紧密连锁的基因的功能丢失所致。一些患 AS 的儿童具有两个 *PWS* 基因的功能性拷贝，反之亦然，但这种过表达似乎不具有任何表型效应。

如表所示，多种事件均可导致相关 15 号染色体序列上父源（PWS）或母源（AS）拷贝的缺失。

- 15q11-q13 旁侧的重复序列之间 4～4.5Mb 的突变型缺失是最常见的原因。父源染色体上的缺失导致 PWS，母源染色体上相同区域的缺失将导致 AS。
- **单亲二体**（uniparental disomy）发现于 DNA 标记研究显示具有外观正常的染色体的个体，自双亲之一继承了特定的一对同源体（此病例中为 15 号），常见原因为**三体拯救**（trisomy rescue）。一个 15 号三体孕体发展到某个多细胞阶段时，一般会死亡，但如果一个偶然发生的有丝分裂不分离（节 2.5.2）在足够早发育阶段产生仅具有两条 15 号染色体拷贝的细胞的话，该细胞能够继续发育，产生存活下来的婴儿。大多数 15 号三体孕体为 $15^M15^M15^P$，在 1/3 的情况下，一条 15 号染色体的随机丢失，将产生母源性单亲二体的胎儿，15^M15^M。而任一父源 15 号染色体的缺乏，胎儿都会患上 Prader-Willi 综合征。

框 16.6 Prader-Willi 和 Angelman 综合征的分子病理学（续）

- 偶尔，印迹产生的机制将发生错误。两条同源 15 号染色体，将具有相同的亲代特异性甲基化模式，尽管标记研究提示它们来自不同的亲代。这些有趣的例子是由决定的非重叠性的 PWS 和 AS **印迹调控元件**（imprinting control element）的小缺失所引起的。
- AS 可能完全是由 *UBE3A* 基因的表达缺乏所致，因为一些遗传性病例具有正常的染色体结构和印迹，但在这个基因中有点突变。PWS 则更为复杂。父源染色体编码一个巨大的（多至 460kb 及 148 个外显子）具有多种剪接体的转录物。前 10 个外显子编码两个蛋白，SNURF 和 SNRPN，而下游的外显子则缺乏读框，但一些内含子编码形成部分剪接机构的核仁小 RNA（snoRNA）。snoRNA 的缺失可能导致 PWS 症状。转录物的下游部分与另一条链上的 *UBE3A* 基因相重叠，可能扮演反义 RNA 的角色，阻止父源染色体上 *UBE3A* 的转录（Runte *et al.*, 2001）。

Prader-Willi 和 Angelman 综合征的起源

事件	PWS 的比例	AS 的比例
缺失	约 75%	约 75%
单亲双体	约 20%	约 3%
印迹错误	约 2%	约 5%
点突变	未见	约 15%（*UBE3A* 基因内）

16.5 功能获得性突变

表 16.3 中列举了几种能够产生功能获得性表型的机制。

表 16.3 功能获得性突变的机制

功能紊乱	基因	疾病	MIM 号
过度表达	*PMP22*	Charcot-Marie-Tooth 病	118200
受体永久“开启”	*GNAS*	McCune-Albright 病	174800
获得新底物（*Pittsburgh* 等位基因）	*PI*	α1 抗胰蛋白酶缺陷	107400
离子通道不正常开放	*SCN4A*	先天性肌强直	168300
结构异常的多聚体	*COL2A1*	成骨不全	各种
蛋白聚集	*HD*	Huntington 病	143100
嵌合基因	*BCR-ABL*	慢性髓样白血病	151410

16.5.1 新功能的获得在遗传性疾病中少见，但常见于癌症中

一个随机改变的基因将更可能终止其工作，而几乎不可能使之产生新的功能。通常产生具有新功能的基因的唯一机制是，染色体重排将两个不同基因的功能性外显子连在一起（图 17.4A）。毫无疑问，这类外显子的重排在进化过程中非常重要；对分子病理学来说，这种改变在导致癌症时最常被注意到。许多获得性的肿瘤特异性染色体重排将产生导致细胞增殖失控的具有新活性的嵌合基因（表 17.3）。遗传性点突变赋予蛋白质新功能的一个罕见例子是 *PI* 基因座上的 *Pittsburgh* 等位基因（MIM 107400；图 16.6）。

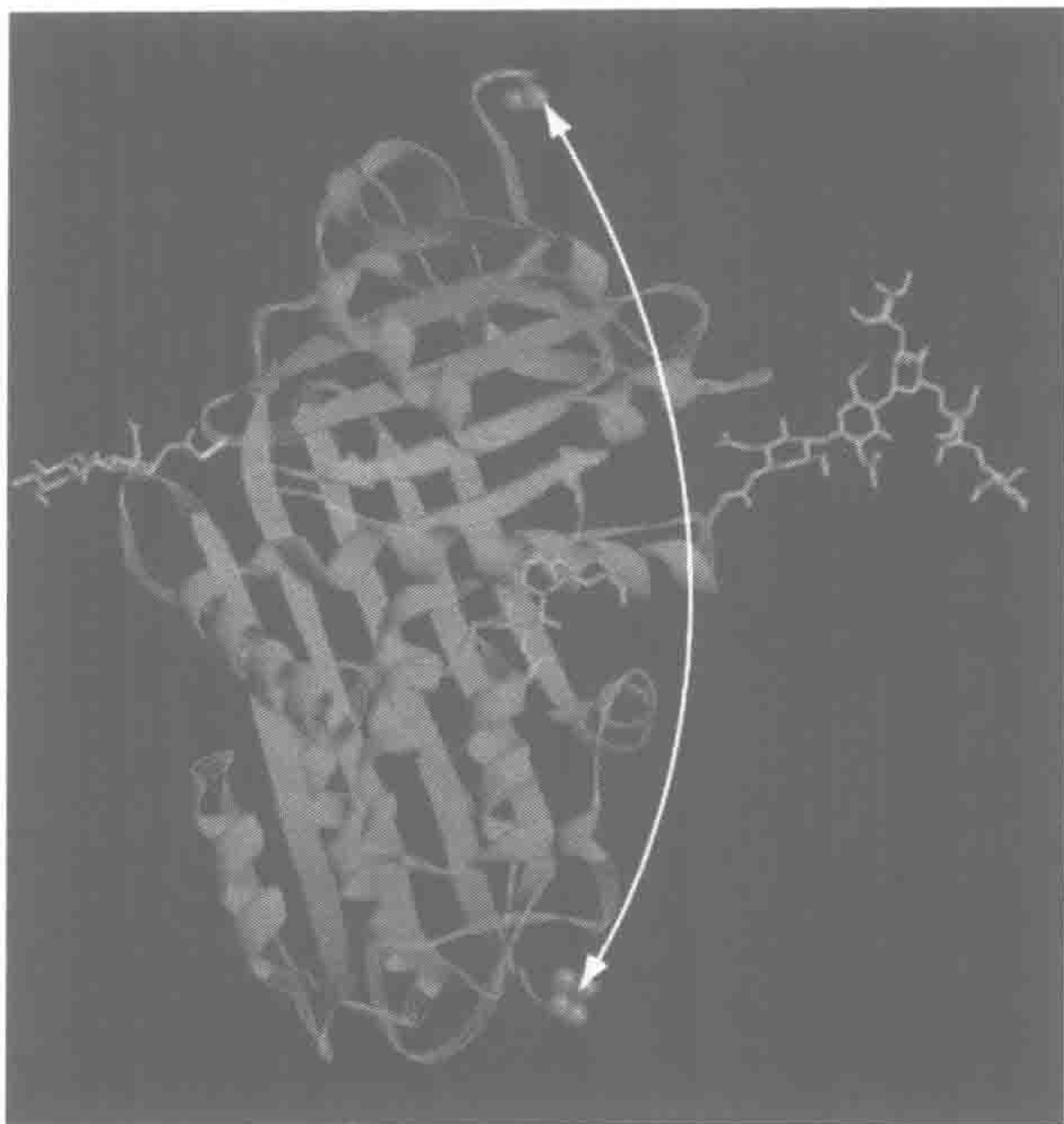


图 16.6 遗传性突变导致蛋白获得新功能

α1 抗胰蛋白酶活性中心的蛋氨酸 358 担当弹性蛋白酶的“诱饵”。弹性蛋白酶切割蛋氨酸 358 和丝氨酸 359 之间的肽链连接，造成这两个残基弹开呈 65Å，如图所示（绿球）。弹性蛋白酶被捕获，并失活。Pittsburgh 变异体含有一个错义突变 M358R，将蛋氨酸诱饵替换为精氨酸。这将破坏其对于弹性纤维酶的亲和性，但产生一个凝血酶的诱饵。作为新的结构性活性抗纤维酶原，Pittsburgh 变异体将导致一种致死性出血病。图像来自于 University of Geneva 的 ExPASy 分子生物学 World Wide Web 服务器。

16.5.2 过度表达可能具有致病性

癌细胞中特定基因的明显过度表达较为常见。体细胞遗传变异导致过度表达的机制包括基因的大量重复扩增，或将正常情况下低水平表达的基因易位至高活性表达的染色质环境中，这些将在节 17.3.3 中充分讨论。

遗传性疾病不常缘于单一基因的构成型高表达。Xp21.3 上 *DSS* 基因的复制将导致男性至女性的性别逆转，可能是剂量加倍的直接后果（Bardoni *et al.*, 1994）。对于 *PMP22* 外周髓鞘蛋白基因而言，基因剂量由 2 个增加至 3 个拷贝即足以导致 Charcot-Marie-Tooth 疾病（见下文及图 16.8）。这类基因表达的中度增强可能很少为致病性，尽管未知基因的类似程度的剂量敏感性肯定可以解释染色体三体的许多特征（节 16.8.2）。异常基因产物的超活性，在基因正常转录及翻译的情况下，能够产生相似的效应。

16.5.3 基因产物的数量改变可以引起功能获得

尽管真正的新功能的获得在遗传性疾病中很少见，改变细胞信号反应的活化突变常产生显性表型。G 蛋白偶联的激素受体提供了一个好例子。许多激素通过结合至跨膜受

体的细胞外结构区域对靶细胞发挥功效。配体的结合将引起受体在细胞质中尾部催化失活的（GDP 结合）G 蛋白转变为活性的（GTP 结合）形式，而这将通过刺激腺苷酸环化酶引起信号进一步传导。有些突变即便在无配体的情况下也能使受体激活腺苷酸环化酶。

- 家族性男性青春期早熟（MIM 176410：受累男孩 4 岁即进入青春期）被发现具有持续激活的黄体激素受体。
- 常染色体显性甲状腺增生可由促甲状腺激素受体的活化突变引起（MIM 275200）。
- Jansen 干骺端软骨发育不良（MIM 156400：一种骨生长异常）构成性的活化甲状旁腺激素受体。
- 构成型激活的 $G_S\alpha$ 蛋白（受体偶联 G 蛋白的一部分）可导致 McCune-Albright 综合征或多骨性纤维发育不良（PFD，MIM 174800）。PFD 仅见于嵌合体的体细胞——可能是因为构成型突变为致死性。根据携带突变细胞系的组织不同，结果可能为多骨性纤维发育不良，浅褐斑，性早熟及其他功能异常增强性内分泌。相同基因的功能丢失性突变常常会导致不同的疾病，即 Albright 遗传性骨营养不良（表 16.5）

16.6　分子病理学：从基因到疾病

思考分子病理学的出发点可能是某个基因或一种疾病。而对于分子病理学的完整理解理应将两者结合起来，这两个路径在本节和下节中将分开来考虑。

16.6.1　对于功能丢失性突变来说，表型的影响将取决于残留的基因功能水平

表 16.1 中描述的 DNA 序列改变可能导致不同程度的功能丢失。许多氨基酸替代具有很小或没有影响，而一些突变将完全破坏功能。突变可能存在于基因的一个或两个拷贝上。患有常染色体隐性疾病的人常常携带有两个不同突变，为**复合杂合子**（compound heterozygote）。如果两个突变均造成功能丢失，而程度不同，较轻的等位基因将决定残留功能的水平。

图 16.7 示意了基因残留功能的水平与临床表型之间的 4 种可能关系。

A. 简单的隐性疾病。倘若杂合个体发生完全破坏基因功能的突变，但剩下的等位基因无明显的缺陷的话，个体表型正常。

B. 由单体性不足所致的显性疾病。事实上，这种简单情况较少见。如果基因产物减少 50% 将导致临床症状，更严重的减少将可能产生更加严重的影响。

C. 严重程度不同的隐性疾病。众多的例子中包括：

- X 连锁的次黄嘌呤鸟嘌呤磷酸核糖转移酶（HPRT）基因的突变。突变体中的残留酶活性程度与受累男性的临床表型对应良好（表 16.4）。
- α 珠蛋白基因拷贝数下降将产生逐次加重的影响。如图 16.2 所示，大部分人具有 4 个拷贝的 α 珠蛋白基因（αα/αα），有 3 个拷贝的人（αα/α－）健康，有 2 个的（不论其状态为 α－/α－或 αα/－）将患轻度 α 地中海贫血，仅有 1 个基因（α－/－）者将患严重疾病，而所有 α 基因（－/－）的缺失将导致致死性胎儿水肿。

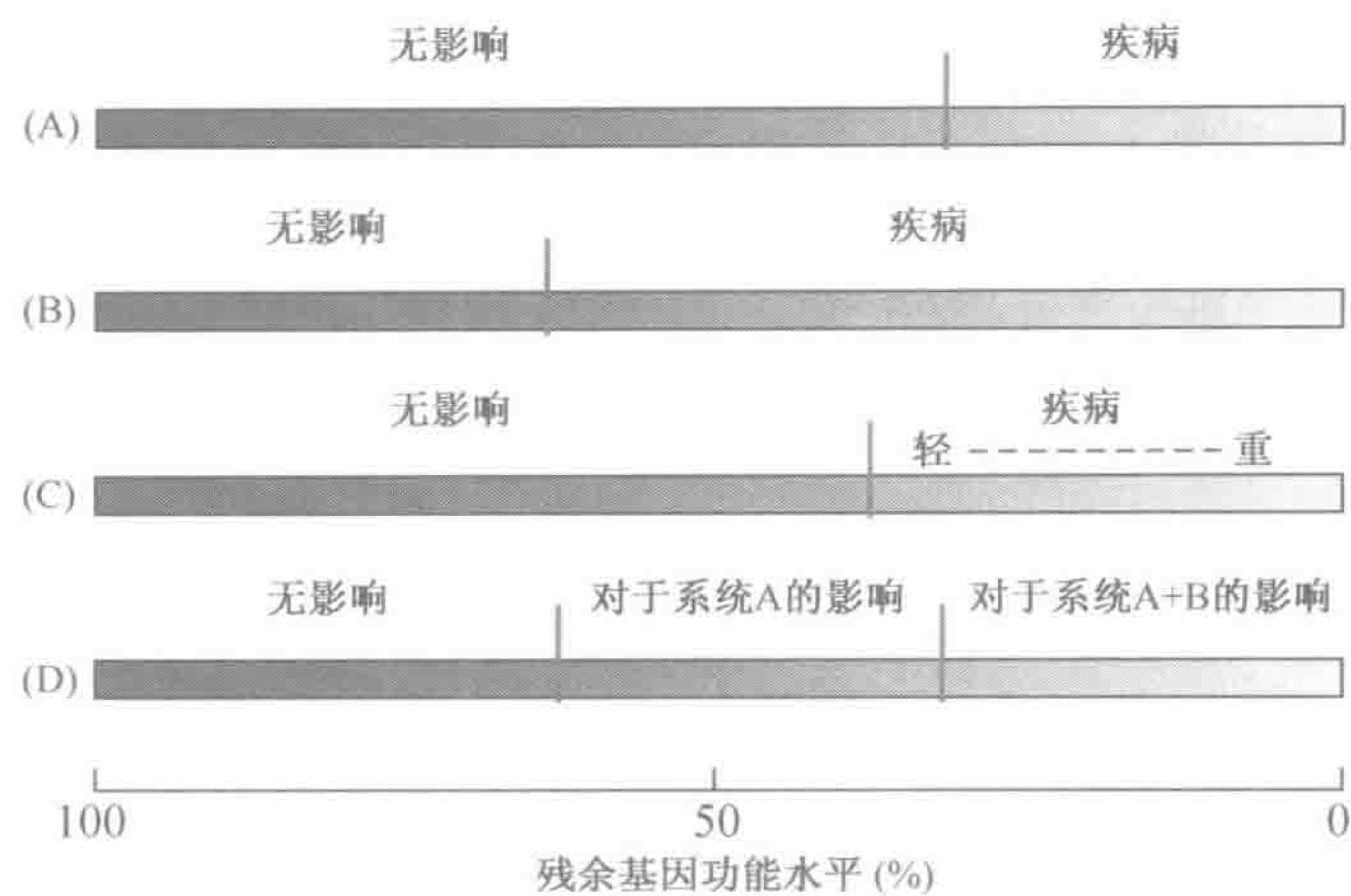


图 16.7　功能丢失与临床表型之间的 4 种可能关系

见节 16.6.1 中的讨论。

表 16.4　次黄嘌呤鸟嘌呤磷酸核糖转移酶功能下降的后果

HRPT 活性(正常的%)	表型
>60	正常
8～60	神经学正常;高尿酸血症(痛风)
1.6～8	神经学异常(舞蹈手足徐动症)
1.4～1.6	Lesch-Nyhan 综合征(舞蹈手足徐动症,自残)但智力正常
<1.4	典型的 Lesch-Nyhan 综合征(MIM 308000,舞蹈手足徐动症,自残及智力低下)

D. 几种表型。一个基因残余功能的降低可能扩延其表型，可能导致具有不同临床标签的疾病。根据阈值的位置，可出现几种不同的情况：

- 几种相关的隐性疾病可由单一基因座上基因功能的连续下降引起。例如细胞外基质富含硫酸蛋白聚糖样硫酸乙酰肝素和硫酸软骨素，而硫酸盐转运的缺陷会干扰骨骼发育。*DTDST* 硫酸盐转运因子的功能丢失性突变，根据其丢失的程度将导致 4 种存在关联的常染色体隐性骨发育不良，即弯曲变形发育不良（MIM 226600)、多发性骨骺发育不良 4（MIM 226900)、骨发育不全症Ⅱ（MIM 256050）及软骨发育不全 1B 型（MIM 600972)。
- 50%的丢失可能没有影响；而由显性负效应引起更显著的丢失则可能导致显性疾病；而功能的完全丢失则将导致更为严重的隐性疾病。*KVLQT1* K^+ 通道的简单功能丢失性突变对杂合子没有影响，但在纯合子中则将导致隐性 Jervell and Lange-Nielsen 综合征（MIM 220400：心脏异常及听力丧失)。然而，同一基因产物的显性负性突变将产生显性遗传的 Romano-Ward 综合征（MIM 192500：心律失常)。在转染的爪蟾卵母细胞中，Romano-Ward 离子通道具有约 20%的正常功能，但在 JLN 患者中离子通道的功能则完全丢失（Wollnik *et al.*,1997)。

- 同一基因的突变可导致两种或更多的显性疾病，由单纯性单倍性不足将导致轻微的异常，显性负效应则将导致更为严重的疾病。这发生在编码Ⅰ型胶原的 *COL1A1* 或 *COL1A2* 基因中（图 16.4）。这些基因中的突变通常将导致成骨不全（OI：脆骨病）。移码和无义突变将导致 1 型 OI，即最轻微的类型，而 Gly-X-Y 重复单元中的氨基酸替换则见于更为严重的Ⅱ、Ⅲ、及Ⅳ型 OI 中。基因型与表型之间的关系很微妙。Gly-X-Y 单位中的甘氨酸被更大的氨基酸替代将通过破坏胶原三螺旋的紧密包装而呈现显性负效应。三螺旋自 C 端开始包装，靠近该末端的甘氨酸替代比 N 端附近被替代的影响更为严重。外显子 6（*COL1A1* 或 *COL1A2*）被越过具有大为不同的影响。N 端原肽裂解位点丢失以及异常胶原的产生可导致 Ehlers-Danlos 综合征Ⅶ型（MIM 130060；皮肤及关节松弛）。丢失了多种不同的功能，从而产生了表型。

16.6.2　同一基因的功能丢失与功能获得性突变将导致不同疾病

我们已经看到，*PAX3* 基因功能丢失性突变将导致发育异常 1 型 Waardenburg 综合征（图 16.1）。在体细胞中，一个获得的染色体易位使 *PAX3* 与另一个转录因子基因 *FKHR* 融合而产生一个新的嵌合基因时，可见到完全不同的表型。这种杂交转录因子的功能获得将引起儿童肺泡横纹肌肉瘤的发生（表 17.3）。

一个惊人的例子涉及 *RET* 基因（Manié *et al.*,2001），*RET* 编码一个跨细胞膜的受体。当其配体（GDNF）与细胞膜外结构域结合时，它将诱导受体二聚化，后者随后通过它们的细胞质结构域中的酪氨酸激酶将信号传递到细胞中。多种功能丢失性突变——阻碍 RET 蛋白翻译后成熟的移码、无义突变以及氨基酸替代——是 Hirschsprung 病的原因之一（MIM 142623，大肠中肠神经结缺如所致的顽固性便秘，节 15.6.2）。*RET* 基因的某些非常特殊的错义突变可见于一组截然不同的疾病，即家族性甲状腺髓质瘤以及相关但更为广泛的多发性内分泌瘤 2 型中。这些属于功能获得性突变，产生对配体过度反应，或者是即便在配体缺失的情况下也能持续激活和二聚化的受体。奇怪的是，影响到 618 或 620 位半胱氨酸的部分携带错义突变的患者同时患有甲状腺癌及 Hirschsprung 病——同时发生的功能获得和丢失。这提醒我们，功能获得与功能丢失并不总是单纯的标量；在存在基因表达的不同类型细胞中，突变可能具有不同的效应。

单一基因的突变可能导致一种以上的疾病，表 16.5 中列举了几个例子。通常功能获得性突变体产生质量异常的蛋白质。偶尔单纯的剂量效应可能引起致病性——外周神经髓鞘蛋白基因 *PMP22* 就是一个例子。染色体 17p11 上的重复序列间的不等交换将产生包括 *PMP22* 基因在内 1.5Mb 区域的重复或缺失（图 16.8）。重复或缺失的杂合子携带者分别具有该基因的 1 个或 3 个拷贝。仅有单一拷贝者将患遗传性神经病伴发压迫性麻痹神经病或香肠样神经病变（MIM 162500），而如上所述，具有 3 个拷贝者将出现一种不同的临床表现的神经病变：Charcot-Marie-Tooth 病 1A（CMT1A；MIM 118220）。

表 16.5　导致一种以上疾病的基因实例

基因	位置	疾病	符号	MIM 号
PAX3	2q35	Waardenburg 综合征 1 型	WS1	193500
		肺泡横纹肌肉瘤	RMS2	268220
CFTR	7p31.2	囊性纤维化	CF	219700
		双侧输精管缺如		
RET	10q11.2	多发性内分泌瘤 2A 型	MEN2A	171400
		多发性内分泌瘤 2B 型	MEN2B	162300
		甲状腺髓质瘤	FMTC	155420
		Hirschsprung 病	HSCR	142623
PMP22	17p11.2	Charcot-Marie-Tooth 神经病变 1A 型	CMT1A	118220
		香肠样神经疾病	HNPP	162500
SCN4A	17q23.1-q25.3	先天性肌强直	PMC	168300
		高钾性周期性瘫痪	HYPP	170500
		先天性乙酰唑胺反应性肌强直		
PRNP	20p12-pter	Creutzfeldt-Jakob 病	CJD	123400
		家族性致死性失眠	FFI	176640
GNAS	20q13.2	Albright 遗传性骨营养不良	AHO	103580
		McCune-Albright 综合征	PFD	174800
AR	Xcen-q22	睾丸女性化综合征	TFM	313700
		脊延髓肌萎缩症	SBMA	313200

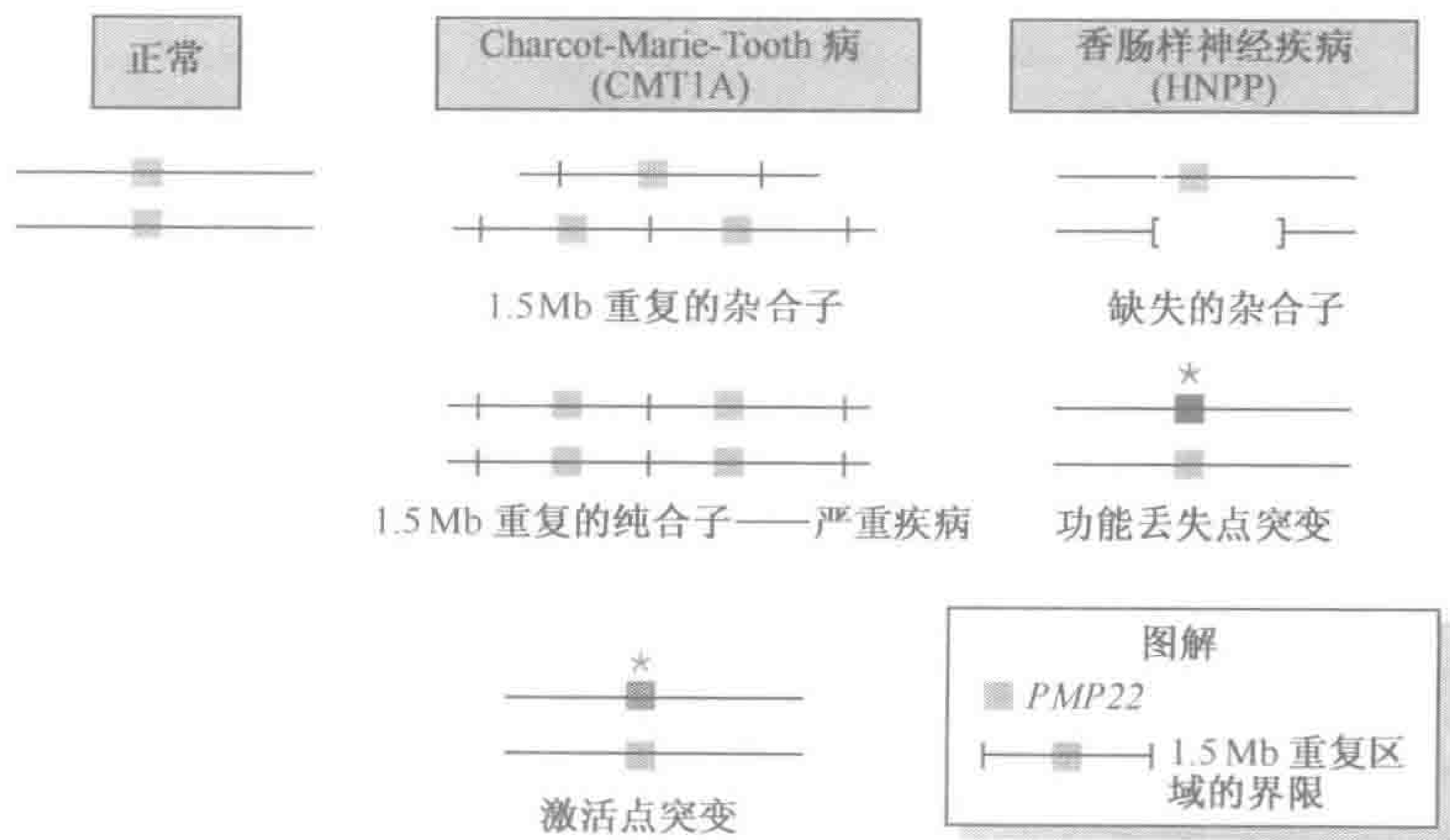


图 16.8　*PMP22* 基因的基因效量效应

大多数 Charcot-Marie-Tooth 病的患者为 17p11.2 包括外周髓鞘蛋白基因，*PMP22* 处 1.5 Mb 重复的杂合子，纯合子重复具有非常严重的疾病，部分患者仅有两个拷贝的 *PMP22* 基因，但其中一个拷贝带有一个活化的突变。*PMP22* 基因的缺失或功能丢失性突变可见于香肠样神经病变中（Patel and Lupski，1994）。

16.6.3　家系中的差异是修饰基因或偶然效应的证据

许多孟德尔疾病即便在携带完全一样突变的同一家系的受累成员之间也表现出临床差异。家系内的差异肯定是由其他不相关基因（修饰基因）以及环境效应（包括偶然事件）的某种联合效应所致。如前所述，单倍性不足的表型对修饰因素的影响尤为敏感（节 16.4.2）。Waardenburg 综合征是一个典型的例子：图 16.1 显示，这种疾病是由单

倍性不足引起的证据，而图 4.5C 显示了典型的家系内差异。

在遗传咨询中，家系内的可变性是一个大问题，因为计划要孩子的家庭想知道孩子受累的情况将有多重。因此，存在临床以及科学的动机去识别修饰基因。候选的修饰基因可以通过有关与主要的基因产物的生化作用的知识获得，或者来自对于可进行必要遗传分析的小鼠的研究（Nadeau，2001）。Easton 等（1993）对于神经纤维瘤 I 型的研究示范了对于大家系内临床表型的统计分析将如何提供修饰基因存在的证据。纯粹偶然事件的角色也不应该忽略，尤其是在具有杂凑表型的疾病中。实例包括 Waardenburg 综合征中的斑状色素减退，以及分别见于神经纤维瘤 I 型与腺瘤样结肠息肉（MIM 175100）中的数目不等的神经纤维瘤或息肉。

修饰因子如何作用的一个实例来源于一个具有明显双基因遗传的眼白化病的有趣家系（Morell *et al.*,1997）。带有明显的两性遗传的眼病家庭是修饰基因因子如何作用的实例（Morell *et al.*,1997）。酪氨酸酶是黑色素细胞的一个关键酶；缺陷将导致眼皮肤白化病（MIM 203100），酪氨酸酶基因的一个常见变异体 R402Q 编码活性降低的酶，残存的活性已足够高，能使纯合子拥有正常的色素。然而，在 Morell 等所报道的家系中，当他们同时携带 1 或 2 个 R402Q 拷贝时就表现出眼白化病（眼皮肤白化病的较轻类型），涉及黑色素细胞分化的 *MITF* 基因也发生突变。*MITF* 突变不能独自导致眼白化病。

16.6.4　不稳定扩延的重复——一类新的病因

1991 年首次发现时，不稳定扩张性核苷酸重复（**动态突变**，dynamic mutation）是一种全新而无前例的疾病机制。目前已知的实例如表 16.6 所示。它们提出了两个重大问题：

- 不稳定性和扩延的机制是什么？这已在节 11.5.2 中得到了讨论。
- 扩张的重复为什么会使你生病？这将在此讨论。

由动态重复所致的疾病的一个标志是**早现遗传**（anticipation）——即在后续的世代中，发病年龄更小和/或程度更为严重。有关早现遗传的提示见节 4.3.3。在某些情况下，中等长度的等位基因为非致病性但不稳定，并易于扩延为完全的突变等位基因（例如，50～200 单位的 FRAXA 重复）；在另一些情况下，这类等位基因仅非常罕见地发生扩延（例如，有 29～35 个重复的 *HD* 等位基因）。致病性扩延可分为两类：

- 编码序列外极度扩延的重复；
- 编码在基因产物中多聚谷氨酰胺束的 CAG 重复的中度扩延。

编码序列外极度扩延的重复

在脆性 X 综合征和 Friedreich 共济失调中，庞大的扩延重复通过破坏转录而导致功能丢失。在青少年肌阵痉挛癫痫中扩延的 12 聚体也是如此。在每个例子中，疾病偶尔是由于该基因中不同的、更常规的功能丢失性突变所致。除可能不表现早现遗传外，这类突变将导致与扩延相同的临床表型。其他类似的极度扩延重复，诸如 *FRA16A*（扩延的 CCG 重复）或者 *FRA16B*（扩延的 33bp 小卫星）为非致病性，可能是附近没有重要的基因。重复很可能通过改变局部染色质结构而破坏转录。在脆性 X 综合征中，这造成了启动子的甲基化。

表 16.6　由核苷酸重复不稳定扩延所致的疾病

疾病	MIM 号	遗传方式	基因位置	重复位置	重复序列	稳定重复数	不稳定重复数
1. 编码序列外重复单位非常大的扩延							
脆性 X 位点 A(FRAXA)	309550	X	Xq27.3	5′非翻译区	(CGG)*n*	6～54	200～1000+
脆性 X 位点 E(FRAXE)	309548	X	Xq28	启动子	(CCG)*n*	6～25	200+
Friedreich 共济失调(FA)	229300	AR	9q13-q21.1	内含子 1	(GAA)*n*	7～22	200～1700
强直性肌营养不良 1 (DM1)	160900	AD	19q13	3′非翻译区	(CTG)*n*	5～35	50～4000
强直性肌营养不良 2 (DM2)	602668	AD	3q21	内含子 1	(CCTG)*n*	12	75～11000
脊髓小脑共济失调 8 (SCA8)	603680	AD	13q31	未翻译 RNA	(CTG)*n*	16～37	110～500+
脊髓小脑共济失调 10 (SCA10)	603516	AD	22q13	内含子 9	(ATTCT)*n*	10～22	Up to 22kb
青少年肌阵挛癫痫 (JME)	254800	AR	21q22.3	启动子	(CCCCGCCCCGCG)*n*	2～3	40～80
2. 编码序列中 CAG 重复的中度扩延							
Huntington 病(HD)	143100	AD	4p16.3	编码	(CAG)*n*	6～35	36～100+
Kennedy 病	313200	XR	Xq21	编码	(CAG)*n*	9～35	38～62
脊髓小脑共济失调 1 (SCA1)	164400	AD	6p23	编码	(CAG)*n*	6～38	39～83
脊髓小脑共济失调 2 (SCA2)	183090	AD	12q24	编码	(CAG)*n*	14～31	32～77
Machado-Joseph 病	109150	AD	14q32.1	编码	(CAG)*n*	12～39	62～86
脊髓小脑共济失调 6 (SCA6)	183086	AD	19p13	编码	(CAG)*n*	4～17	21～30
脊髓小脑共济失调 7 (SCA7)	164500	AD	3p12-p21.1	编码	(CAG)*n*	7～35	37～200
脊髓小脑共济失调 17 (SCA17)	607136	AD	6q27	编码	(CAG)*n*	25～42	47～63
齿状核红核(DRPLA) 苍白球丘脑下部核萎缩	125370	AD	12p	编码	(CAG)*n*	3～35	49～88

肌强直性营养不良中的巨大扩延则有所不同。在肌强直性营养不良患者中从未发现过其他突变，因此 CTG 重复肯定具有某种特殊的地方。最近已澄清，其主要的致病作用（在 DM1 和 DM2 中）是通过 mRMA 结合以及隔离其他转录物正常剪接所需的 CUG 结合蛋白。在其他转录物中，异常的剪接将导致肌肉特异性氯离子通道的丢失（Tapscott and Thornton，2001 综述）。由于扩延的部位形成了一个邻近基因 *SIX5*（MIM 600963）C_pG 岛的一部分，亦可能存在其他效应，扩延将降低该基因的表达。

SCA8 基因（Koob *et al*.，1999）似乎编码一段未翻译的 RNA，作为另一条 DNA 链上的一个基因的反义调节因子（与节 16.4.4 中提及的印迹转录物相似）。扩延如何导致疾病尚属未知。的确，还不能完全确定这种改变为致病性。

编码区内编码基因产物中多聚谷氨酰胺束的CAG重复的中度扩延

由基因内不稳定性CAG重复扩延所致的8种疾病的共同特征包括：

- 它们均为晚发性神经退行性疾病，除Kennedy疾病外，均呈显性遗传；
- 在基因中尚未发现其他突变可以导致该病；
- 扩延的等位基因被转录并翻译；
- 三核苷酸重复编码蛋白质中的一个多聚谷氨酰胺束；
- 存在一个关键的阈值重复大小，低于它的重复为非致病的，而高于它的则可能导致疾病；
- 超过阈值之上，重复越大，平均发病年龄越早（对单个患者无法进行预测，但存在明确的统计相关性）。

Kennedy病中的雄激素受体突变提供了明确的证据，即CAG重复相关疾病涉及一种特殊的功能获得。该基因功能丢失性突变为人所知，并能导致雄激素不敏感或睾丸女性化综合征（MIM 300068），一种男性性别分化的障碍。相反，多聚谷氨酰胺的扩延，将导致一种非常不同的神经退行性疾病，尽管患者也常表现轻微的女性化。常见的致病机制涉及蛋白质聚集的形成（下面）。

除了经典的不稳定扩延重复外，一些疾病可由更短得多的相当稳定的三核苷酸的重复扩延所致。例子包括口咽肌营养不良（*PABP2* 基因，MIM 602279）、假性软骨发育不全（*COMP* 基因，MIM 600310）以及多指（*HOXD13* 基因，MIM 186000）。这些并非动态突变，也不属于典型的不稳定扩延重复。

扩延重复的实验室诊断

用一次PCR反应就可对多聚谷氨酰胺重复疾病进行诊断。肌强直性营养不良中的巨大扩延需要用Southern印迹杂交。脆性X可根据*FRAXA*基因的大小以及甲基化状态Southern印迹杂交法诊断。详细内容及凝胶图片见节18.4.3及图18.12。

16.6.5　在功能获得性疾病中，蛋白质聚集是常见的致病机制

最近，变得明确的是，蛋白质聚集的形成是若干成年发病的神经性疾病的一个共同特征，包括前面描述的多聚谷氨酰胺病、Alzheimer病、Parkinson病、Creutzfeld-Jakob病以及被称为淀粉样变性的一组混杂病变。完整的情况还不清楚，但共同的主题正显现出来。球状的蛋白分子就像重油滴，疏水基在内部，极性基在外部。正确的折叠是一个至关重要而高度特异性的过程，自然形成的蛋白质是从所有可能的蛋白质序列中选择出来，部分程度上由于它们能够正确折叠。突变的蛋白质更容易错误折叠。具有暴露的疏水基团的错误折叠可以相互或与其他蛋白聚集，似乎对神经元以及可能其他细胞具有毒性。某些时候，似乎一种构象变化能够在一群蛋白分子中传播，通过一种与晶体化类似的过程将它们从稳定、天然的构象转变为具有不同性质的新形态。朊蛋白是最显著的例子。错误的折叠可能源自一个新合成的正常结构分子的偶然错误折叠（散发型病例）、更趋于错误折叠的一段突变序列（遗传型病例）、或一个以某种方式获得于环境的错误折叠的分子（感染型病例）。因此，这种最后的共同通路将起因非常不同的一系列

疾病集中到了一起（Perutz and Windle，2001；Bucciantini *et al*.,2002）。

16.6.6 对于线粒体突变，异质性和不稳定性使基因型和表型之间的关系复杂化

可破坏紧密包装的线粒体基因组上的基因功能的点突变、缺失或重复（图 9.2），与涉及中枢神经系统，心脏、肌肉、内分泌系统、肾脏和肝脏在内的广泛的退行性疾病相关。细胞内含有许多 mtDNA 分子。它们可能是同质（每个 mtDNA 分子都是一样的）或异质的（正常与突变的 mtDNA 的混合群体）。与嵌合现象不同的是，异质性可以通过异质的卵子从母亲传递给子女。在个体中突变和异质性常常似乎随时间而变化。同一个体可以同时携带缺失和重复，并且比例可随时间而变化（Poulton *et al*.,1993）。

同样的序列变化也常见于患不同综合征的人群中，因而表型-基因型的相关性特别难以建立。例如，50％的患 Leber 遗传性视神经病（MIM 535000：突发性不可逆视觉丧失）的人在线粒体基因组核苷酸 11778 位存在 G→A 的替换。大多数此类患者为同质性，但约 14％同样严重的受累者为异质性。即便在同质性家系中，情况也是高度可变的；外显率在整体上为 33％～60％，82％的受累个体为男性（Wallace *et al*.,2001）。这种不佳的对应的原因可能包括：

- 异质性可以为组织特异性，被检验的组织（通常为血液或肌肉）可能并不是发病过程中的关键组织；
- mtDNA 比核 DNA 的可变性要大得多，一些症状也许缘于已报道的突变和其他未知的变异体的联合作用；
- 一些线粒体病似乎属于数量性状：小的突变性变异不断积累而降低线粒体产生能量的能力，达到某个阈值时临床症状即出现；
- 线粒体的许多功能由核基因编码（框 9.2），核内的变异因而可能是线粒体表型的重要原因或修饰因子。

线粒体突变的 MITOMAP 数据库（http://www.mitomap.org）中有不错的一般讨论，加上广泛的数据表，恰好显示了预测表型的难题是多么地重要。

16.7 分子病理学：从疾病到基因

在很多时候思考分子病理学的出发点是疾病而不是基因。这种策略为基因型-表型的对应提供了另一种视角。总的启示就是当预测导致一个孟德尔综合征的基因缺陷时，一定不要想得太天真。

16.7.1 疾病的致病基因可能并非显而易见

导致一种蛋白质缺陷的突变并不一定就在编码该蛋白质的结构基因内部

丙种球蛋白缺乏症（缺乏免疫球蛋白，造成临床免疫缺陷）通常为孟德尔型的。对疾病的自然的推测就是免疫球蛋白基因内的突变。但免疫球蛋白基因定位于 2、14 和 22 号染色体，而丙种球蛋白缺乏症并不定位在这些位置。许多类型为 X 连锁。回想起将一个新合成的多肽转变为正常工作的蛋白质所需要的许多步骤（节 1.5），这种突变

和蛋白结构基因之间缺乏一一对应也并不值得大惊小怪。在免疫球蛋白基因加工、B细胞成熟，或者免疫系统的整体发育过程中的故障都将会导致免疫缺陷。

一个基因缺陷有时可以造成多个酶的缺陷

Ⅰ细胞病或黏膜脂肪沉积病Ⅱ型（MIM 252500）的特征为多种溶酶体酶的缺陷。原始的缺陷并不在于这些酶的结构基因上，而是在一个N乙酰葡萄糖胺-1磷酸转移酶上。后者可磷酸化糖基化酶分子上的甘露糖残基。磷酸甘露糖是将酶定位到溶酶体上的信号；当它缺乏时，溶酶体就将缺乏整个酶系。

突变常常仅影响一部分组织的表达基因

基因的组织特异性表达模式是突变的临床效应的较差预测者。基因不表达的组织不太可能发生主要的病变，但反过来就不对。通常仅有一部分表达组织受累。*HD* 基因广泛地表达，但 Huntington 病仅累及大脑的局限区域。视网膜母细胞瘤基因（节17.6.1）普遍地表达，但遗传性突变通常仅累及视网膜。这在溶酶体疾病中也同样令人吃惊。基因表达在许多组织中的单一类型细胞的巨噬细胞中的表达是必需的，但在受累患者中并非所有包含巨噬细胞的组织都是异常的。其解释并不难发现：

- 基因并非仅表达于需要它们的组织中。只要表达无害，可能并没有选择压力来关闭表达，即使表达在并不带来益处的组织中；
- 基因功能的丢失对一些组织的影响比其他组织大得多，因为不同的细胞类型的不同作用和代谢需求，以及细胞内相互作用的网络中功能冗余程度的不同。来自于癌遗传学的“看门人基因”的概念（节17.7.2）可能也适用于许多其他细胞功能和机能失常，以及肿瘤细胞中失常的细胞更新；
- 任何功能获得均可能对某些细胞类型为致病性而对其他细胞类型则无害，见 *RET* 基因的例子（节16.6.2）。

16.7.2 基因座异质性是惯例而非例外

基因座异质性（locus heterogeneity）是指同一疾病可由几个不同基因的突变所致的情况。重要的是应该思考基因产物的生物学功能以及与其相互作用的分子，而不是期望基因和综合征之间的一一对应的关系。正如在节4.2.4所见，临床综合征常常缘于发育或生理通路的故障或机能失常所致；同样，许多细胞的结构和功能依赖于多成分的蛋白聚集。如果需要几个基因的正常工作，那么其中任何基因的突变都可能导致相同或非常相似的表型。

胶原蛋白（图16.4；节16.6.1）再次提供了一个好例子。我们已经知道Ⅰ型胶原，即皮肤、骨骼、肌腱和韧带的主要胶原，是由两条α(1)链和一条α(2)链的三螺旋构成。*COL1A1* 或 *COL1A2* 基因的突变将导致相同的疾病，即显性成骨不全。Ⅱ型胶原形成软骨以及包括眼玻璃体在内的其他组织中的纤维。它是由COL2A1链的同源三聚螺旋组成。*COL2A1* 基因的不同突变导致包括 Stickler 综合征、脊柱骨骺发育不良及 Kniest 发育不良在内的范围重叠的一组骨发育不良。相似的表型可由Ⅺ型胶原中的突变所致，它是Ⅱ型纤维中较少的组分。在所有这些病例中，将产生何种综合征将取决于最终胶原

纤维的总体效应，而非哪个基因发生了突变。

16.7.3 基因家族中不同成员的突变可产生一系列相关或重叠的综合征

具有部分重叠功能的基因家族中成员的突变可产生一组部分重叠临床上难以分开的表型。影响成纤维细胞生长因子受体的突变说明了这一点。10 种成纤维细胞生长因子通过 4 种细胞表面受体 FGFR1～4 调控重要的发育过程。大多数组织表达多种 FGFRS，包括每种的剪接变异体。FGFRS 为受体酪氨酸激酶，以相似的方式与前面描述的 RET 蛋白相互作用，信号转导有赖于受体的二聚化，而后者可能涉及同源二聚体或异源二聚体。FGFRS 突变体可能造成剪接体间平衡的改变，改变了同源和异源二聚体之间的平衡，通过显性负效应或产生固有的活性二聚体而减弱信号。因此，存在复杂遗传效应的可能。

受体基因非常特殊的突变将导致一系列骨生长的显性疾病（图 16.9）。10q26 上 FGFR2 的突变可见于 Crouzon、Jackson-Weiss、Pfeiffer 以及 Apert 综合征，而 4p16 上 FGFR3 的另外的特异突变则将导致软骨发育不全、致死性发育不良 1 型及 2 型、软骨发育不良、伴有黑棘皮症的 Crouzon 综合征以及 Muencke 冠状颅缝早闭。一些 Pfeiffer 综合征患者具有 FGFR1 突变。这些综合征的临床描述、参考文献和分子病理学的介绍见 OMIM 和 Wilkie（1997）。突变的最特异性本质提示，功能获得、软骨发育不良、致死性发育不良及 Crouzon 突变体已证实，当转染进特定类型细胞中时，可产生不同程度的组成性活性（配体无关）的受体（Naski *et al.*,1996）。

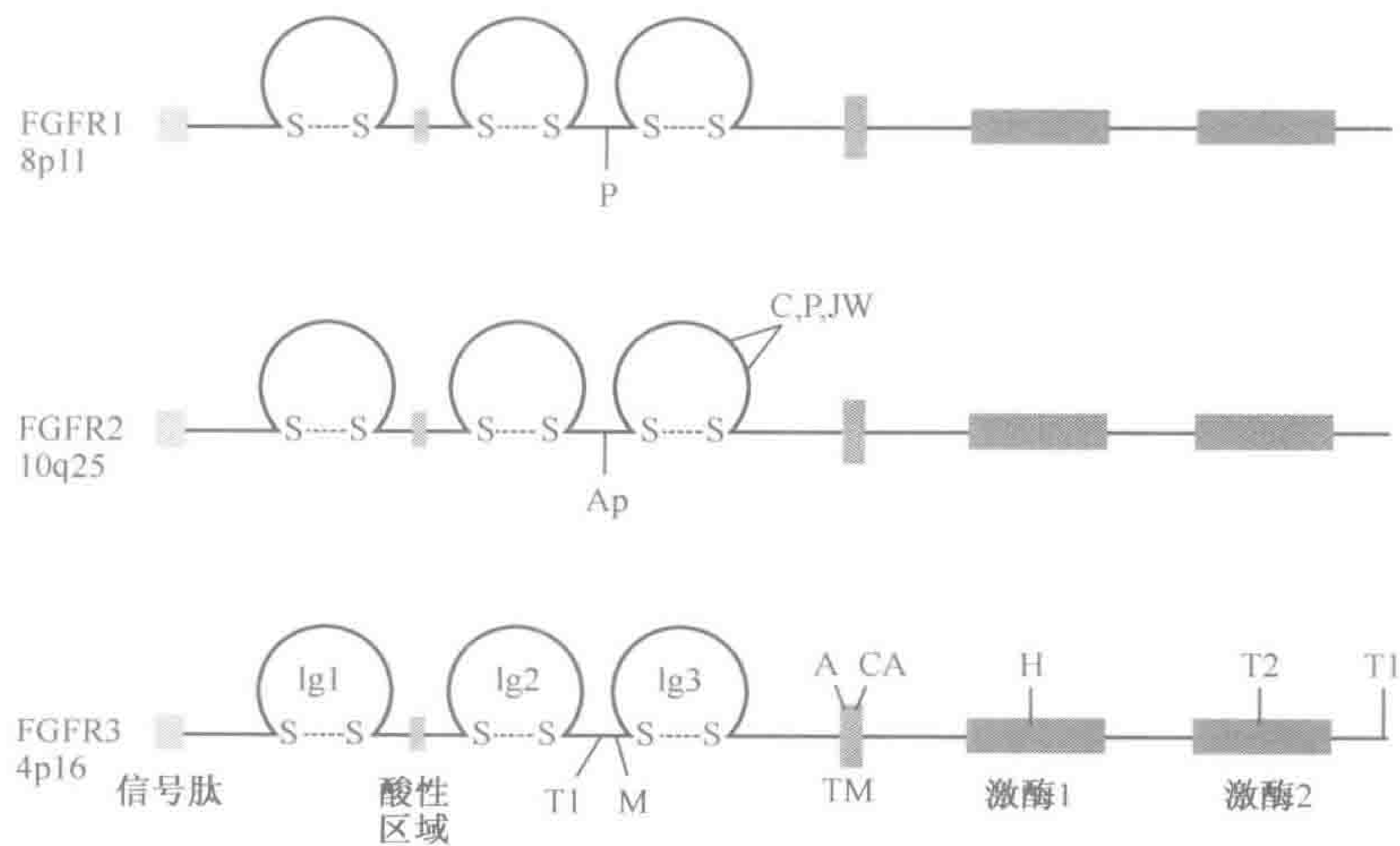


图 16.9 FGFR 突变的表型-基因型相关性

4 个高度同源的成纤维细胞生长因子受体中的 3 个如图所示。每个受体酪氨酸激酶有 3 个免疫球蛋白样细胞外结构域（由 S-S 桥连接）、一个跨膜结构域（TM）以及匹配的细胞内酪氨酸激酶结合域。非常特殊的错义突变与一系列的骨发育不良相关（软骨发育不全 A、软骨发育不良 H、致死性发育不良 1 型及 2 型，T1 和 T2）和颅缝早闭综合征（Apert Ap，Crouzon C，Jackson-Weiss JW，Muencke M，Pfeiffer P），其他突变则可导致 Beare-Stevenson 回状头皮皮肤病（CA）。FGFR2 的 Ig3 结构域的一些突变与 Crouzon、Jackson-Weiss 及 Pfeiffer 综合征不同家系存在关联。

16.7.4 临床与分子分类是思考疾病的不同工具，二者在各自的领域均有其正当性

本章中的一个反复主题，即由胶原蛋白基因突变所致的结缔组织疾病示范了疾病的临床与分子分类之间的差别（表 16.7）。

▶ 所有的孟德尔疾病均在分子基础上进行分类，首先按所涉及的基因座，然后按该基因座上特定的突变等位基因。

▶ 遗传病也可根据它们的症状和预后进行临床分类。这种方式定义的临床分类也许并不能与分子分类完全对应，但它们对于患者的咨询和治疗可能更有用。

表 16.7A　在正文及表 16.7B 中提及的结缔组织病的临床分类

疾病	MIM 号	特征
成骨不全(脆骨病)Ⅰ型(根据牙科所见分为ⅠA、ⅠB、ⅠC)	166200 (166240)	轻至中度骨脆性；蓝色巩膜；正常身材；听力丧失(50%)
成骨不全(脆骨病)Ⅱ型	166210	非常严重的骨脆性；围产期致死性
成骨不全(脆骨病)Ⅲ型	(166230)	中至重度骨脆性；渐进性变形；极矮身材；常见听力丧失
成骨不全(脆骨病)Ⅳ型(根据牙科所见分为ⅣA、ⅣB)	166220	轻至中度骨脆性；正常巩膜；各种身材
脊柱骨骺发育不良	183900	短小身材；短颈；脊柱骨骺发育不良
Stickler 综合征	108300	轻度脊柱骨骺发育不良；腭裂；高度近视；听力丧失
Kniest 发育不良	156550	不成比例的短小身材；短颈；脊柱骨骺发育不良等
Ehlers-Danlos 综合征Ⅶ型	130060	关节及皮肤松弛

表 16.7B　在正文及表 16.7A 中提及的结缔组织病的分子分类

基因	位点	突变	综合征
COL1A1	17q22	无效等位基因	成骨不全Ⅰ型
		部分缺失；C 端替代	成骨不全Ⅱ型
		N 端替代	成骨不全Ⅰ、Ⅲ或Ⅳ型
		剪接突变；外显子 6 缺失	Ehlers-Danlos 综合征Ⅶ型
COL1A2	7q22.1	剪接突变；外显子缺失	成骨不全Ⅰ型
		C 端突变	成骨不全Ⅱ、Ⅳ型
		N 端替代	成骨不全Ⅲ型
		外显子 6 缺失	Ehlers-Danlos 综合征Ⅶ型
COL2A1	12q13	点突变	脊柱骨骺发育不良
		无义突变	Stickler 综合征
		转换缺陷	Kniest 发育不良
		无义突变	软骨发生不全Ⅱ型
COL11A2	6q21.3	剪接突变	Stickler 综合征

临床标记并非简单的规范。它们随着基础遗传知识的进展而发展——疾病被分为同类（Duchenne 和 Becker 肌营养不良）或分开（散发型乳腺癌与 BRCA1 乳腺癌）。分子分类对分子诊断至关重要，并可以提供更加精确的咨询。例如，只有分子诊断才能够证明，生育有一个以上成骨不全患儿的非受累双亲为性腺嵌合体而非隐性类型 OI 的携带者。然而，成熟的分子分类对临床并非总是有用。例如，尽管 OMIM 列出了 11 个可导

致 Usher 综合征（隐性聋哑）的位点，临床上，它仅对区分严重程度不同的三种类型有用。因此，分子分类可以阐明但不能替代临床分类。

16.8　染色体病的分子病理学

16.8.1　微缺失综合征填补了单个基因与染色体综合征之间的空白

倘若我们的 3000 Mb 基因组包含 30 000 个基因，一个 100 万碱基左右的缺失，尽管小得在显微镜下都难以看见，仍可能涉及十几个乃至于更多的基因。FISH 分析和微阵列 CGH（节 2.4.2，17.3.3）正在揭示越来越多的这类亚显微水平染色体畸变（表 16.8）。从分子病理学角度看，它们可分为三类。

表 16.8　染色体微缺失相关综合征

由单一基因单倍性不足所致的综合征常见于并无微缺失，但却具有基因内点突变的患者中

综合征	MIM 号	位点	异常类型
Wolf-Hirschhorn	194190	4p16.3	节段非整倍性
Cri du chat	123450	5p15.2-p15.3	节段非整倍性
Williams	194050	7q11.23	节段非整倍性
Langer-Giedon	150230	8q24	邻接的基因(*TRPS1*,*EXT1*)
WAGR	194072	11p13	邻接的基因(*PAX6*,*WT1*)
Prader-Willi	176270	15q11-q13	节段非整倍性(缺失父源性拷贝)
Angelman	105830	15q11-q13	缺失母源性 *UBE3A*
Rubinstein-Taybi	180849	16p13.3	*CBP* 单倍性不足
Miller-Dieker	247200	17p13.3	邻接的基因(*LiS1*,*YWHAE* 等)
Smith-Magenis	182290	17p11.2	节段非整倍性
Alagille	118450	20p12.1	*JAG1* 单倍性不足
Di George/VCFS	192430	22q11.21	节段非整倍性

WAGR，Wilms 瘤，无虹膜，生殖器异常，智力障碍；VCFS，腭-心-面综合征。

- **单基因综合征**（single gene syndrome），所有的表型效应均缘于单一基因的缺失（或有时为重复）。例如，Alagille 综合征（MIM 118450）可见于携带一个 20p11 微缺失的患者中。然而，93%的 Alagille 患者并无缺失，而是具有位于 20p11 上的 *JAG1* 基因的点突变。在所有病例中该综合征的原因是 *JAG1* 的单倍性不足。
- **邻接基因综合征**（contiguous gene syndrome），主要见于携带 X 染色体缺失的男性。经典的病例为"BB"男孩，患有 Duchenne 肌营养不良（MIM 310200）、慢性肉芽肿病（MIM 306400）和色素性视网膜炎（MIM 312600）以及智力低下（Francke *et al.*，1985）。他具有一个 Xp21 染色体缺失，后者除去了一组邻接的基因，附带为研究者提供了克隆这些基因的方法，其缺失即导致其中两种疾病，即 DMD 与慢性肉芽肿病（节 14.4.2）。Xp 末端的缺失可见于另一组邻接基因综合征中。逐渐加大的缺失将除去更多的基因并向综合征中增加更多的疾病（Ballabio and Andria，1992）。微缺失在 X 染色体的一些部位（如 Xp21，Xq 近侧）中相对常见，但在其他部分（如 Xp22.1-22.2，Xq28）中则少见或未知。毫无疑问，特定的个别基因的缺失，以及在基因富集区内的可见缺失，将会是致死性的。

▶ **节段非整倍性综合征** (segmental aneuploidy syndrome)，常染色体微缺失很少属于真正的邻接基因综合征。通常杂合子的表型仅取决于缺失基因中的一部分，即剂量敏感者（图 16.10）。若干已阐释清楚的综合征是由再发性的突变型微缺失所致（Budarf and Emanuel，1997）。易于缺失的区域被长的重复所包围，后者将容许错配性重组（图 11.17）。这些重复常含有转录序列，可形成更加开放、易于发生重组的染色体结构。这些百万碱基大小区域中的一部分倒置为常见的非致病性多态，后者可能通过使旁侧重复错配而使缺失易于发生（Giglio *et al*.,2001）。

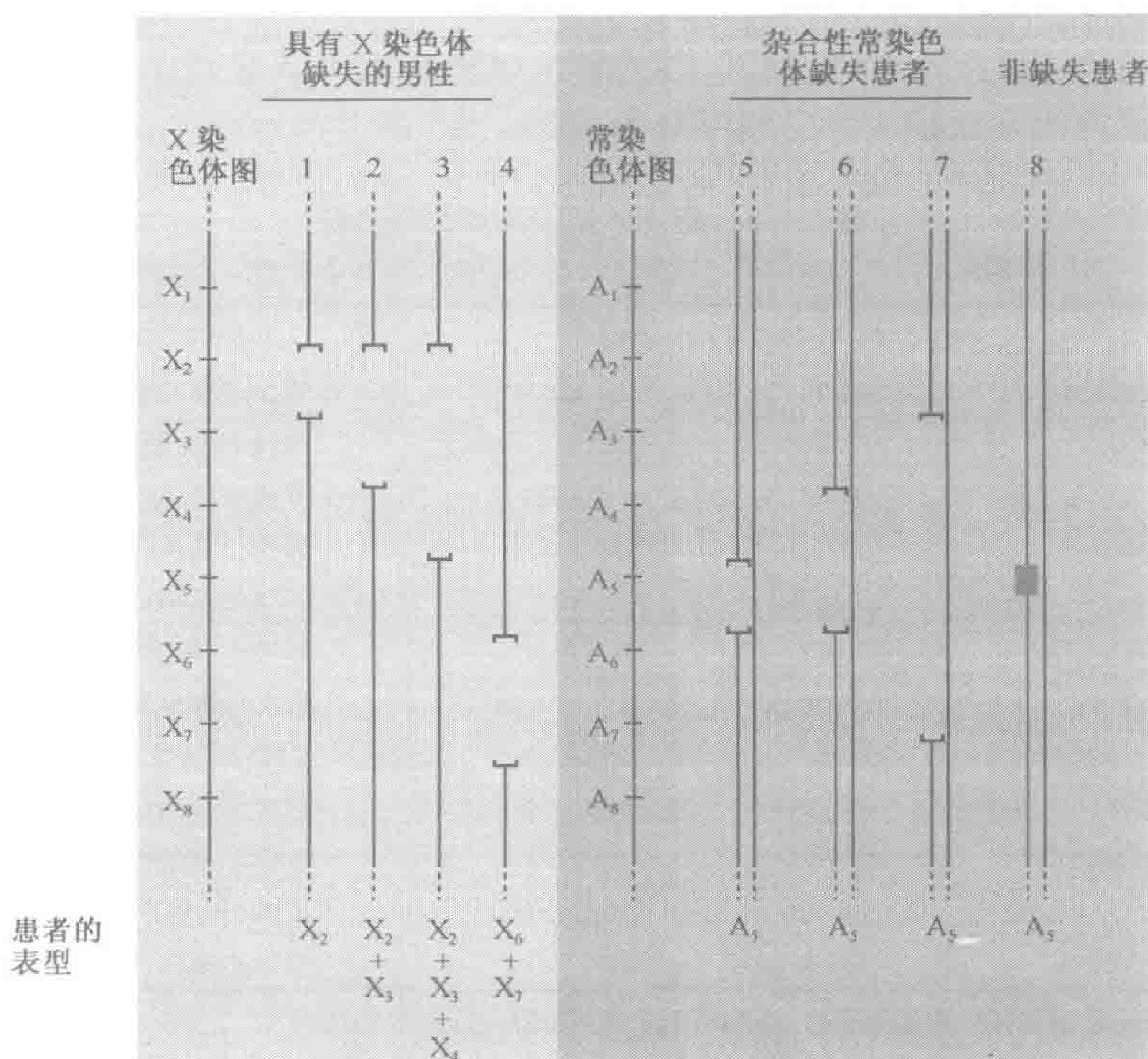


图 16.10　X连锁和常染色体微缺失综合征

在X染色体上，基因 X_1 或 X_5 的缺失在男性中为致死性。患者1～3呈现一系列的嵌套邻接基因综合征，而患者4则具有不同的邻近基因综合征。在常染色体上，仅基因 A_5 为剂量敏感。携带有不同缺失的患者5～7呈现与患者8相同的表型，后者为 A_5 基因功能丢失性点突变的杂合子。

识别缺失区域中哪些基因导致了微缺失综合征中的哪些表型被证明是困难的。三种可能的策略为：

▶ 寻找患有由候选区域中单一基因的突变引起、按孟德尔方式遗传的综合征的组分之一的患者；

▶ 寻找携带比正常小的微缺失，并且仅表现综合征的部分特征的个体；

▶ 删除小鼠中的相应区域；将缺失小鼠和该区域个别基因的转基因小鼠进行杂交，将表型与剩余的单体性不足的程度相对应。

Williams 综合征（WLS，MIM 194050）为这些问题提供了一个好例子。患 WLS 的人具有可识别的面容，他们生长迟滞，他们在婴儿期可能具有威胁生命的低血钙，并

且他们经常具有主动脉瓣膜狭窄（SVAS）。此外，他们通常为中度智力低下，与患Down综合征的人大致相同，但他们有非常特殊的认知特征与个性。他们非常好交际，通常喜欢音乐，谈吐异常地好，但在使用形状（视觉空间构建能力）上却具有特殊的残疾。WLS是由一个1.6 Mb的缺失所致，为旁侧重复重组的结果，尤其是在具有整个区域的一个常见倒位的杂合型个体中（Osborne *et al.*,2001）。从缺失区域已鉴定出大约20个基因（Tassabehji *et al.*,1999）。在所有缺失于WLS者中鉴定相关的基因可能会为发现正常的人类认知与行为的遗传决定因子提供一条途径。

如前所述（节16.4.1），SVAS可由弹性蛋白基因的缺失或破坏所致。该基因位于Williams关键区内，而弹性蛋白的单倍性不足毫无疑问是见于Williams综合征中的SVAS的病因。面部特征也可能是由这种结缔组织蛋白的缺乏所致，然而这显然并非事实，因为具有简单弹性蛋白突变的人常有SVAS但却没有Williams面容。该综合征中尚无其他特征被令人信服地归因于这一区域内的其他任何基因。具有该区域部分缺失的人似乎要么只有SVAS，要么具有完全的综合征。Williams表型是否能在小鼠中被识别是个有趣的问题。其他的微缺失综合征亦与特异性的行为相关，因此阐明这些综合征的组成基因是目前一个很有兴趣的研究。

16.8.2 染色体非整倍性的主要效应可能是由几个可辨认的基因的剂量失衡所致

单体和三体所具有的特征性表型很可能缘于重叠于发育过程中许多微小紊乱之上的几个主要基因的效应。剂量增加50%即可产生某种重大效应的基因一定很不寻常，因此，应有可能发现引起例如Down综合征（DS）主要特征的少数基因。对携带易位的患者的研究表明，DS关键区域位于21q22.2；21号染色体其他部分的三体并不导致DS。至少有两个DS特征的候选基因已在该区域内被发现：*DYRK*，其果蝇和小鼠同源体（*minibrain*）将导致剂量敏感性学习障碍（Altafaj *et al.*,2001），以及*DSCAM*，一种表达于发育过程神经系统和心脏中的细胞黏附分子（Barlow *et al.*,2001）。

X染色体单体和三体尤其有趣，因为X失活应该使它们在身体组织中无症状。然而，并非所有的X连锁基因均失活。Turner综合征的骨骼肌异常是由*SHOX*单体性剂量不足所致，后者为一个Xp/Yp假常染色体区域内的同源框基因（节2.3.3，图12.15）（Clement-Jones *et al.*,2000）。其他身体特征则很可能缘于逃避了X失活且具有功能性Y对应物的其他基因的单体性不足（图12.14）。已知的这类基因仅有18个（Lahn and Page，1997），因此潜在候选者的清单似乎还可以对付。

（吕晶玉 译）

进一步阅读

Dipple KM, McCabe ERB (2000) Phenotypes of patients with 'simple' Mendelian disorders are complex traits: thresholds, modifiers and system dynamics. *Am. J. Hum. Genet.* **66**, 1729–1735.

Scriver CR, Waters PJ (1999) Monogenic traits are not simple: lessons from phenylketonuria. *Trends Genet.* **15**, 267–272.

Weatherall DJ (2001) Phenotype–genotype relationships in monogenic disease: lessons from the thalassaemias. *Nature Rev. Genet.* **2**, 245–255.

Weatherall DJ, Clegg JB, Higgs DR, Wood WG (2001) The hemoglobinopathies. In: *The Metabolic and Molecular Basis of Inherited Disease*, 8th Edn (eds CR Scriver, AL Beaudet, WS Sly, D Valle). McGraw Hill, New York, pp. 4571–4636.

参考文献

Altafaj X, Dierssen M, Baamonde C *et al.* (2001) Neurodevelopmental delay, motor abnormalities and cognitive deficits in transgenic mice overexpressing *Dyrk1A* (*minibrain*), a murine model of Down's syndrome. *Hum. Molec. Genet.* **10**, 1915–1923.

Ballabio A, Andria G (1992) Deletions and translocations involving the distal short arm of the human X chromosome: review and hypotheses. *Hum. Molec. Genet.* **1**, 221–722.

Bardoni B, Zanaria E, Guioli S *et al.* (1994) A dosage sensitive locus at chromosome Xp21 is involved in male to female sex reversal. *Nat. Genet.* **7**, 497–501.

Barlow GM, Chen X-N, Shi ZY *et al.* (2001) Down syndrome congenital heart disease: a narrowed region and a candidate gene. *Genet. Med.* **3**, 91–101.

Bird AC (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21.

Bucciantini M, Giannoni E, Chiti F *et al.* (2002) Inherent cytotoxicity of polypeptide aggregates suggests a common origin of protein misfolding diseases. *Nature* **416**, 507–511.

Budarf ML, Emanuel BS (1997) Progress in the autosomal segmental aneusomy syndromes (SASs): single or multi-locus disorders? *Hum. Molec. Genet.* **10**, 1657–1665.

Cartegni L, Krainer AR (2002) Disruption of an SF2/ASF dependent exonic splicing enhancer in *SMN2* causes SMA in the absence of *SMN1*. *Nature Genet.* **30**, 377–384.

Clement-Jones M, Schiller S, Rao E *et al.* (2000) The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. *Hum. Molec. Genet.* **9**, 695–702.

Den Dunnen JT, Antonarakis SE (2001) Nomenclature for the description of human sequence variations. *Hum. Genet.* **109**, 121–124.

Easton DF, Ponder MA, Huson SM, Ponder BAJ (1993) An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): evidence for modifying genes. *Am. J. Hum. Genet.* **53**, 305–313.

Francke U, Ochs HD, de Martinville B *et al.* (1985) Minor Xp21 chromosome deletion in a male associated with expression of Duchenne muscular dystrophy, chronic granulomatous disease, retinitis pigmentosa and McLeod syndrome. *Am. J. Hum. Genet.* **37**, 250–267.

Giglio S, Broman KW, Matsumoto N *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms and common chromosome rearrangements. *Am. J. Hum. Genet.* **68**, 874–883.

Hemesath TJ, Steingrimsson E, McGill G *et al.* (1994) Microphthalmia, a critical factor in melanocyte development, defines a discrete transcription factor family. *Genes Dev.* **8**, 2770–2780.

Hentze MW, Kulozik AE (1999) A perfect message: RNA surveillance and nonsense-mediated decay. *Cell* **96**, 307–310.

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature Rev. Genet.* **3**, 415–428.

Karniski LP (2001) Mutations in the diastrophic dysplasia sulfate transporter (DTDST) gene: correlation between sulfate transport activity and chondrodysplasia phenotype. *Hum. Molec. Genet.* **10**, 1485–1490.

Koob MD, Moseley ML, Schut LJ *et al.* (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature Genet.* **21**, 379–384.

Lahn BT, Page DC (1997) Functional coherence of the human Y chromosome. *Science* **278**, 675–680.

Lykke-Andersen J, Shu M-D, Steitz JA (2001) Communication of the position of exon–exon junctions to the mRNA surveillance machinery by the protein RNPS1. *Science* **293**, 1836–1839 (see also the preceding paper in that issue).

Mani, S, Santoro M, Fusco A, Billaud M (2001) The RET receptor: function in development and dysfunction in congenital malformation. *Trends Genet.* **17**, 580–589.

Morell R, Spritz RA, Ho L *et al.* (1997) Apparent digenic inheritance of Waardenburg syndrome type 2 (WS2) and autosomal recessive ocular albinism (AROA). *Hum. Mol. Genet.* **6**, 659–664.

Motulsky AG (1995) Jewish diseases and origins. *Nature Genet.* **9**, 99–101.

Nadeau JH (2001) Modifier genes in mice and men. *Nature Rev. Genet.* **2**, 165–174.

Naski MC, Wang Q, Xu J, Ornitz DM (1996) Graded activation of fibroblast growth factor receptor 3 by mutations causing achondroplasia and thanatophoric dysplasia. *Nature Genet.* **13**, 233–237.

Nicholls RD, Knepper JL (2001) Genome organization, function and imprinting in Prader-Willi and Angelman syndromes. *Annu. Rev. Genomics Hum. Genet.* **2**, 153–175.

Nissim-Rafinia M, Kerem B (2002) Splicing regulation as a potential genetic modifier. *Trends Genet.* **18**, 123–127.

Osborne LR, Li M, Pober B *et al.* (2001) A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nature Genet.* **29**, 321–325.

Patel PI, Lupski JR (1994) Charcot-Marie-Tooth disease: a new paradigm for the mechanism of inherited disease. *Trends Genet.* **10**, 128–133.

Perutz MF, Windle AH (2001) Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats. *Nature* **412**, 143–144.

Poulton J, Deadman ME, Bindoff L, Morten K, Land J, Brown G (1993) Families of mtDNA re-arrangements can be detected in patients with mtDNA deletions: duplications may be a transient intermediate form. *Hum. Mol. Genet.* **2**, 23–30.

Prusiner SB, Scott MR, DeArmond SJ, Cohen FE (1998) Prion protein biology. *Cell* **93**, 337–348.

Reik W, Walter J (2001) Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* **2**, 21–32.

Rougeulle C, Heard E (2002) Antisense RNA in imprinting: spreading silence through Air. *Trends Genet.* **18**, 434–437.

Runte M, Huttenhofer A, Gross S, Kiefmann M, Horsthemke B, Buiting K (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense transcript for *UBE3A*. *Hum. Mol. Genet.* **10**, 2687–2700.

Tapscott SJ, Thornton CA (2001) Reconstructing myotonic dystrophy. *Science* **293**, 816–817.

Tassabehji M, Metcalfe K, Karmiloff-Smith A *et al.* (1999) Williams syndrome: using chromosomal microdeletions as a tool to dissect cognitive and physical phenotypes. *Am. J. Hum. Genet.* **64**, 118–125.

Wallace DC, Lott MT, Brown MD, Kerstann K (2001) Mitochondria and neuro-ophthalmologic diseases. In: *The Metabolic and Molecular Bases of Inherited Disease*, 8th Edn (eds CR Scriver, AL Beaudet, WS Sly, D Valle) McGraw Hill, New York, pp. 2425–2509.

Wilkie AO (1997). Craniosynostosis: genes and mechanisms. *Hum. Mol. Genet.* **6**, 1647–1666.

Wollnik B, Schroeder BC, Kubisch C, Esperer HD, Wieacker H, Jentsch T (1997) Pathophysiological mechanisms of dominant and recessive *KVLQT1* K^+ channel mutations found in inherited cardiac arrhythmias. *Hum. Mol. Genet.* **6**, 1943–1949.

第17章　癌遗传学

本章内容

17.1　前言

肿瘤并非是严格意义上的一种任何多细胞有机体自然终结状态的疾病。我们都熟知这样一个基本的达尔文理论，即在繁殖能力上具有可遗传差异的一群有机体将通过自然选择来进化。繁殖更快或更广泛的基因型将在随后的世代中占据优势，只会又被更高效的繁殖者所替代。其决定因素可能在于出生率的提高或死亡率的降低。同样的道理也适用于构成像人这样的多细胞有机体的细胞群。细胞的生与死受遗传控制。如果体细胞突变产生了一个增殖更快的变异体，那么这个突变克隆就会趋于控制该有机体。因此可以将癌症视为一个自然进化的过程。

癌是一系列体细胞突变的结果，在某些情况下亦具有遗传背景。根据组织来源和显微镜下的组织学类型可以对它们进行分类（癌来源于上皮细胞，白血病和淋巴瘤来源于血液细胞前体等）（图 17.1）。最近，体现肿瘤基因表达总体情况的表达阵列的应用有望使分类进一步细化（图 17.18）。这些分类对决定预后和治疗是重要的，但是，它们并不能解释癌是如何进化的。癌遗传学的目的就是要了解使一个正常的体细胞产生一群不断增殖的侵袭性癌细胞的多步骤突变与选择的途径。

癌遗传学可能非常令人迷惑。每个肿瘤都是独特的。如此之多的不同基因在这种或那种肿瘤中获得突变，并且它们以如此复杂的方式相互作用，使人容易迷失于大量的细节中。在本章中，我们尽量避免混淆，将注意力集中在目前所了解的肿瘤发生的原理上，而不是基因和突变的分类上。

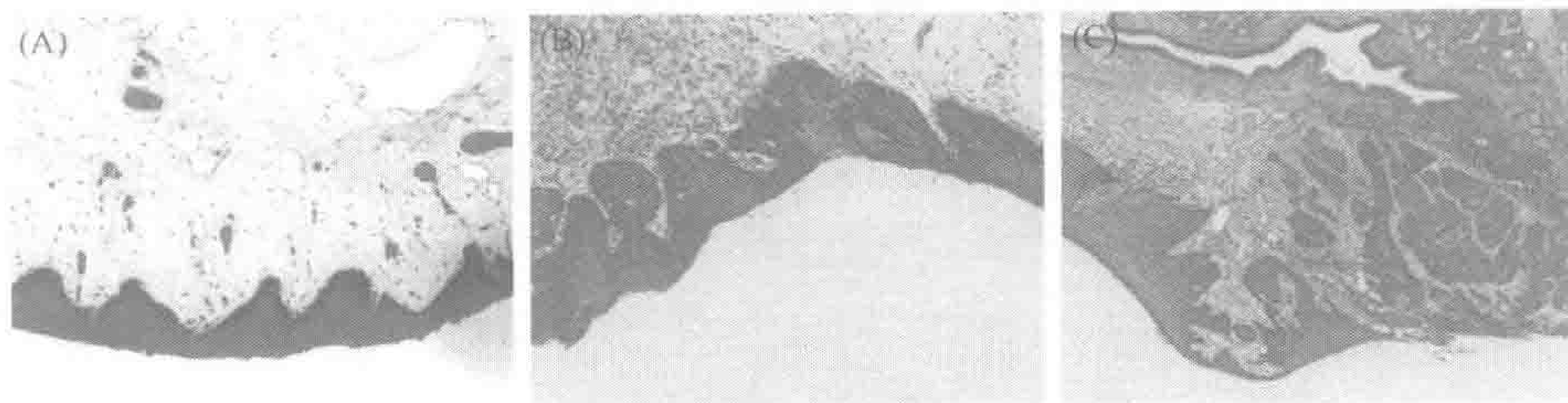


图 17.1　肿瘤的组织学

癌发生的阶段。三张苏木素-伊红染色的口腔黏膜组织切片显示口腔癌发生的阶段。(A) 正常上皮。(B) 发育不良的上皮细胞，是一种潜在的癌前病变。这种上皮细胞表现出紊乱的生长、突变、异常细胞和增多的有丝分裂。(C) 起源于表层上皮并侵入深层结缔组织的癌。肿瘤（癌）岛表现出紊乱的分化、异常细胞、增多且非典型的有丝分裂。病理学家应用这样的组织结构的改变来判定肿瘤并将其分级。由 Nalin Thakker 博士惠赠。

17.2　癌的演化

如上所述，细胞是在强大的选择压力下演变成为肿瘤细胞的。但是，当作为细胞时肿瘤是很成功的，而作为有机体时，它们则是无希望的失败者。在其宿主的寿命之后，它们并不留下任何后代。因此，对整个有机体而言，至少在其生育并抚养大她的子女之前，存在一个强有力的选择机制以防止其死于肿瘤。我们因此被两组相反的选择力量所支配。但是，对肿瘤发生的选择是短期的，而对抵抗力的选择则是长期的。从一个正常的体细胞到一个恶性肿瘤的演变可以发生在个体的一生中，并在一个新的个体中又重新开始。但是，一个具有良好抗肿瘤机制的有机体可将其传给它的后代，使其在那里继续演化。至少在我们的可生育期，10 亿年的演化已经赋予我们高度发达的、连锁重叠的机制以保护我们不患肿瘤。潜在的肿瘤细胞或者被修复而变回正常细胞，或者被迫杀死自身（细胞凋亡）。单一突变不能绕过这些防御机制而使一个正常细胞转变成一个恶性细胞。很久以前，对癌与年龄依赖性的研究曾提示由一个正常上皮细胞转变成具有侵袭力的癌平均需要 6～7 个连续的突变。换言之，只有当 6 个独立防御体系被突变所破坏时，才能将一个正常细胞转变成恶性肿瘤（图 17.2）。

一个细胞经历 6 个独立突变的概率常被人们所忽略，提示癌症应该像即将消失一样

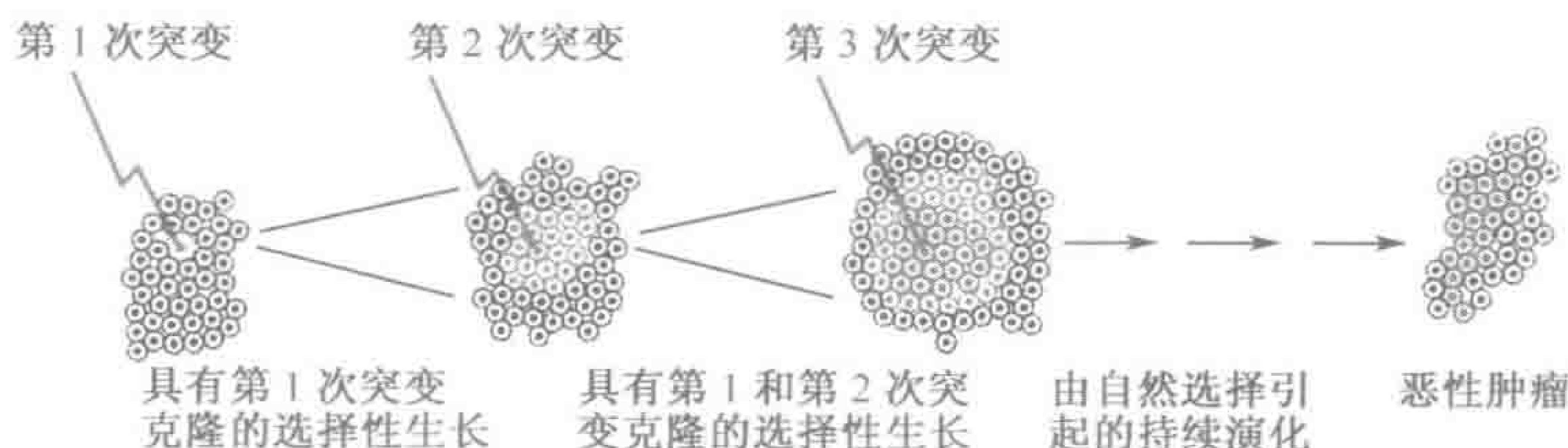


图 17.2　癌的多阶段演变

每次连续的突变均赋予细胞生长优势，进而形成扩大的克隆，为下一次突变提供了更大的靶标。

非常罕见。然而，存在着两种普遍的机制可以促进肿瘤进展的发生（框 17.1）。累积所有的突变无论怎样都需要时间，因此，癌症主要是一种生殖后期的疾病，在这个时期没有选择压力来进一步改善防御能力。

框 17.1 使一系列连续突变更可能发生的两种途径

一个正常上皮细胞变成一个恶性的癌细胞可能需要 6 个特异突变。如果突变率通常为每代细胞每个基因 10^{-7}，那么任何一个细胞可遭受这么多的突变几乎是不可能的（这就是我们大多数人还活着的原因）。一个个体的 10^{13} 细胞中任何一个细胞发生这种情况的概率是 $10^{13}\times10^{-42}$ 或 10^{-29}。无论怎样，癌的发生是由于两种机制的联合作用引起的。

▶ 一些突变增强了细胞的增殖，为下一次突变提供了更大的靶细胞群（图 17.2）；

▶ 一些突变或者在 DNA 水平或者在染色体水平影响整个的基因组，从而增加了总体的突变率。

因为依赖于这两种机制，癌的发生具有阶段性并始于组织增生或良性的生长，而恶性肿瘤细胞由于存在异常核型通常表现出基因组不稳定性（图 17.10）。

鉴于基因是这些突变的靶标，可以将基因分为两大类以示区别，尽管总是按生物学进行分类，但是，与严密的分类方法相比，它们是更有用的认识肿瘤的工具。

▶ **癌基因**（oncogene）（节 17.3）这些是正常情况下促进细胞增殖的基因。肿瘤细胞中功能突变的获得会产生过度或不恰当的活跃形式。单一突变等位基因可以影响细胞的表型。未突变的形式被称为原癌基因。

▶ **肿瘤抑制基因** [tumor suppressor (TS) gene]（节 17.4）TS 基因的产物抑制癌变的过程。癌细胞中的突变形式已丧失了它们的功能。有些 TS 基因产物阻止不适当的细胞周期进展，有些控制异常的细胞进入细胞凋亡，而其他一些则通过保证细胞 DNA 的精确复制、修复和分离来维持基因组的稳定并降低突变率。TS 的两个等位基因必须同时失活才能改变细胞的行为。

如果将癌比喻为一辆公共汽车，我们就可以将癌基因想像成油门，将肿瘤抑制基因比作刹车。持续踩压油门（某癌基因的显性功能的获得）或所有刹车失灵（某肿瘤抑制基因的隐性功能的丧失）将导致车子的失控。

17.3 癌基因

17.3.1 癌基因的历史

癌基因于 20 世纪 60 年代被发现，人们意识到一些动物癌症（特别是白血病和淋巴瘤）是由病毒引起的。一些病毒具有相对复杂的 DNA 基因组（SV40 病毒、乳头状病毒），而其他则属于急性转化性反转录病毒，它们拥有非常简单的 RNA 基因组。一个标准的反转录病毒基因组只有 3 个转录单位：*gag*，编码内部蛋白；*pol*，编码聚合酶；以及 *env*，编码外壳蛋白（每一个转录物裂解后编码若干蛋白）。令人非常兴奋的是，人们发现急性转化性反转录病毒的转化特性完全归因于它们拥有一个额外的基因，即癌基因。短时间里，一些热心者希望所有癌症都可以用携带癌基因的病毒感染来解释，并且希望，一旦这些病毒被鉴定，它们的作用就可能被阻断。

不久，学者发现病毒癌基因是正常细胞基因——原癌基因的拷贝，这些原癌基因偶然地被整合至反转录病毒颗粒中。由于某种原因，病毒癌基因被激活，使它们能转化被感染的细胞。多数人类癌症不依赖病毒，但是无论怎样，它们自身所拥有的原癌基因已被激活。20 世纪末，人们建立起来了一种检测被激活的癌基因的方法。NIH-3T3 小鼠成纤维细胞能被来自人类肿瘤的随机 DNA 片段所转染。其中一些细胞被转化，而且通过构建噬菌体基因组文库和筛选人类特异的 *Alu* 重复，具有转化功能的人类 DNA 可从转化物中被分离出来。从这些转化片段中鉴定的癌基因包括许多曾通过病毒研究获得的已知基因。对有关反转录病毒和细胞癌基因的早期研究工作的描述见 Bishop（1983；见进一步阅读）。

17.3.2　癌基因的功能

对癌基因功能的了解始于 1983 年的一个发现，即病毒癌基因 *v-sis*（前缀 v 表示病毒癌基因）来自正常细胞的血小板源性生长因子 B（*PDGFB*）基因。生长因子失控地过度表达是细胞过度增殖的明显原因。目前，许多细胞癌基因（严格地说是原癌基因）（表 17.1）的作用已被阐明。可喜的是，它们可精确地调控预期在癌中发生紊乱的细胞功能。癌基因按其功能可分为 5 大类：

- ▶ 分泌型生长因子（如 *SIS*）；
- ▶ 细胞表面受体（如 *ERBB*，*FMS*）；
- ▶ 细胞内信号转导系统的成分（如 *RAS* 家族，*ABL*）；
- ▶ DNA 结合核蛋白，包括转录因子（如 *MYC*，*JUN*）；
- ▶ 通过细胞周期控制细胞进展的细胞周期素、细胞周期素依赖性激酶和激酶抑制剂的网络的成分（如 *MDM2*）。

如果癌基因被定义为在肿瘤中发生了占优势的激活突变的基因，那么，目前已知癌基因有 100 多个。

表 17.1　病毒和细胞癌基因

病毒病名	病毒癌基因	细胞癌基因	定位	功能
猴肉瘤	*v-sis*	*PDGFB*	22q13.1	血小板源性生长因子 B 亚单位
鸡红白血病	*v-erbb*	*EGFR*	7p12	上皮生长因子受体
McDonough 猫肉瘤	*v-fms*	*CSF1R*	5q33	巨噬细胞克隆刺激因子受体
Harvey 大鼠肉瘤	*v-ras*	*HRAS*	11p15	G-蛋白信号转导成分
Abelson 小鼠白血病	*v-abl*	*ABL*	9q34.1	蛋白质酪氨酸激酶
禽肉瘤 17	*v-jun*	*JUN*	1p32-p31	AP-1 转录因子
禽髓细胞血症	*v-myc*	*MYC*	8q24.1	结合 DNA 的转录因子
小鼠骨肉瘤	*v-fos*	*FOS*	14q24.3-q31	结合 DNA 的转录因子

病毒基因有时标记为 *v-src*，*v-myc* 等。而它们在细胞中的对应物为 *c-src*，*c-myc* 等。正常细胞的 *c-onc* 基因的形式正式称为原癌基因（proto-oncogene）。如今这些区别常被忽略，而简单地将**癌基因**（oncogene）一词用于正常基因，其异常的形式可被描述为被激活的癌基因。

17.3.3　原癌基因的激活

原癌基因被激活的各种途径为进展中的分子病理学提供了一些最好的例子（表17.2）。激活涉及功能的获得。这种激活或者是量的变化（正常基因产物的增加）或者是质的改变（基因突变产生的微小修饰产物或染色体重排所形成的新的嵌合基因产物）。这些变化具有显性遗传特点，而且正常情况下只影响一个等位基因。Blume-Jensen 和 Hunter（2001）举出其他的例子以说明如下的机制。

表 17.2　激活原癌基因的四种途径

激活机制	癌基因	肿瘤
扩增	*ERBB2*	乳腺癌、卵巢癌、胃癌、非小细胞肺癌、结肠癌
	NMYC	神经母细胞瘤
点突变	*HRAS*	膀胱癌、肺癌、结肠癌及黑色素瘤
	KIT	胃肠基质瘤、肥大细胞瘤
染色体重排产生新的嵌合基因	［许多］	表 17.3
易位至具有转录活性的染色质区	*MYC*	Burkitt 淋巴瘤中 t(8;14)将其易位于免疫球蛋白重链基因座

由扩增激活

许多癌细胞包含结构正常的癌基因的多个拷贝。乳腺癌经常出现 *ERBB2* 的扩增，有时发生 *MYC* 的扩增；其相关基因 *NMYC* 的扩增通常见于晚期神经母细胞瘤和横纹肌肉瘤。可以存在数以百计的额外拷贝。它们或者以分离于染色体的成对的染色质小体（双微体），或者以插入到正常的染色体（均质染色区）中的形式存在。因为它们通常包含源自若干不同染色体的序列，因此，所产生的这些遗传事件也许是相当复杂的（Pinkel，1994 综述）。类似的基因扩增可见于暴露于强烈的人工选择体系的细胞中——例如扩增的二氢叶酸还原酶基因存在于经选择可耐受甲氨蝶呤的细胞中。总之，其结果就是基因表达的水平极大地增加了。

用比较基因组杂交可以研究肿瘤中癌基因的扩增（Forozan *et al.*，1997）。这种分析还能显示等位基因丢失的任何区域或非整倍体，这些结果可以指明肿瘤抑制基因（见下文）。比较基因组杂交实验使用配对的正常和肿瘤组织的 DNA 混合物进行杂交。两种 DNA 分别用红色和绿色荧光物质标记，经混合后作为杂交探针。红色和绿色杂交信号的比值能被检测。可以采取以下两种方法进行比较基因组杂交。

- 标准比较基因组杂交（图 17.3A）将混合的样品与载玻片上的正常染色体杂交（荧光原位杂交，节 2.4.3）。沿每条染色体的长度，绘制红色和绿色荧光原位杂交信号比值的图形，通过可信区间的测量选出偏离期望值 1∶1 的区域。根据偏离的方向，就会将肿瘤中等位基因扩增或缺失的区域标记出来。通过这种技术检测到的最小变化是在 3Mb 左右。
- 阵列-比较基因组杂交（图 17.3B）使用微阵列 DNA 代替染色体进行杂交。每一点

上的 DNA 是由已知染色体区的 BAC 克隆的 DOP-PCR（节 5.2.4）制备而成的。这种技术潜在地提供了更高的分辨率，仅受限于阵列上 BAC 克隆的数量。目前，阵列上 BAC 克隆的数量达到 3000 个，分辨率为 1Mb。与传统的比较基因组杂交相比，这种技术还有一个优点，即可以在 DNA 水平而不是染色体区带水平立刻获得任意一个扩增或缺失的序列。

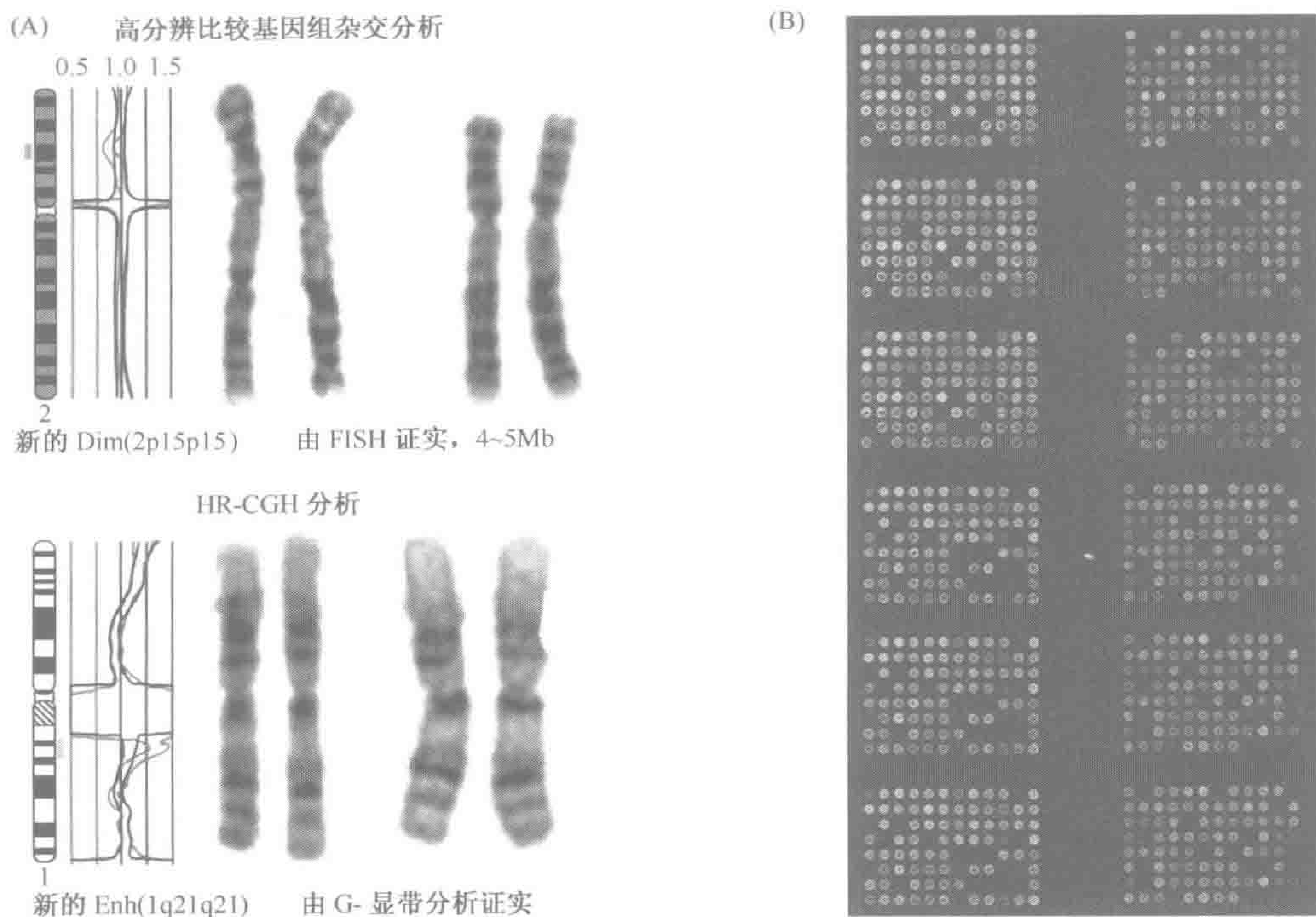


图 17.3　比较基因组杂交

(A) 染色体 *CGH*。两名形态异常和智力低下的男孩样品高分辨比较基因组杂交，他们的 G-显带染色体似乎正常。黄线表示比较基因组杂交的痕迹，黑线表示比值为 1∶1 的 99%可信区间。第一例携带一个新的缺失并经 FISH 证实；第二例携带一个新的扩增区，经回顾，其可见于 G-显带染色体中。图由 Copenhagen 的 Claes Lundsteen 博士惠赠。(B) 阵列 *CGH*。321 个 BAC 克隆的 DOP-PCR 片段被点在载玻片上，然后，与来自乳腺癌细胞系 DNA（绿色标记）和正常参照 DNA（红色标记）的混合物进行杂交。含有肿瘤中扩增 DNA 的 BAC 克隆显示绿色，含有肿瘤中缺失 DNA 的那些 BAC 克隆显示红色，而在肿瘤中无变化的则显示黄色。每个嵌板被点样 3 份以允许杂交效率的变异。由 Nijmegen 的 J. Veltman 博士惠赠。

由点突变激活

三个介导 G-蛋白偶联受体信号（Lowy and Willumsen，1993）的 RAS 家族基因，*HRAS*、*KRAS* 和 *NRAS* 在很多肿瘤中是激活的。配体和这种受体的结合将引起 GTP 和 RAS 蛋白的结合，而 GTP-RAS 则将这种信号在细胞中继续传递下去（图 3.5）。RAS 蛋白具有 GTP 酶活性，GTP-RAS 迅速地转变为无活性的 GDP-RAS。*RAS* 基因的特异性激活点突变经常见于包括结肠、肺、乳腺和膀胱癌在内的多种肿瘤细胞中。突变型 RAS 蛋白具有降低的 GTP 酶活性，以至于 GTP-RAS 失活得更慢，造成细胞对于

受体信号的过度反应。

由产生新嵌合基因的易位激活

这种机制在癌中（上皮肿瘤）很少见，但常见于血液肿瘤和肉瘤中。广为人知的例子是费城（Ph）染色体，这是一种见于 90％慢性髓性白血病患者中的小的近端着丝粒染色体。这个染色体是 9;22 染色体相互平衡易位的产物之一。9 号染色体的断裂点位于 *ABL* 癌基因的一个内含子内。该易位将 *ABL* 基因组 3′端部分序列与 22 号染色体上 *BCR*（断裂点簇集区）基因 5′端部分序列相连，产生了一个新的融合基因（Chissoe *et al.*，1995）。这个嵌合基因表达一种与 *ABL* 产物相关但具有异常转化特性的酪氨酸激酶（图 17.4A）。

目前，已经发现许多肿瘤特异性的断裂点，而且，许多癌基因通过克隆研究已被鉴定（表 17.2）。在癌症基因组解剖学计划的网站上可以查询到一个包含大约 40 000 种染色体畸变的数据库（Mitelman *et al.*，2002）。

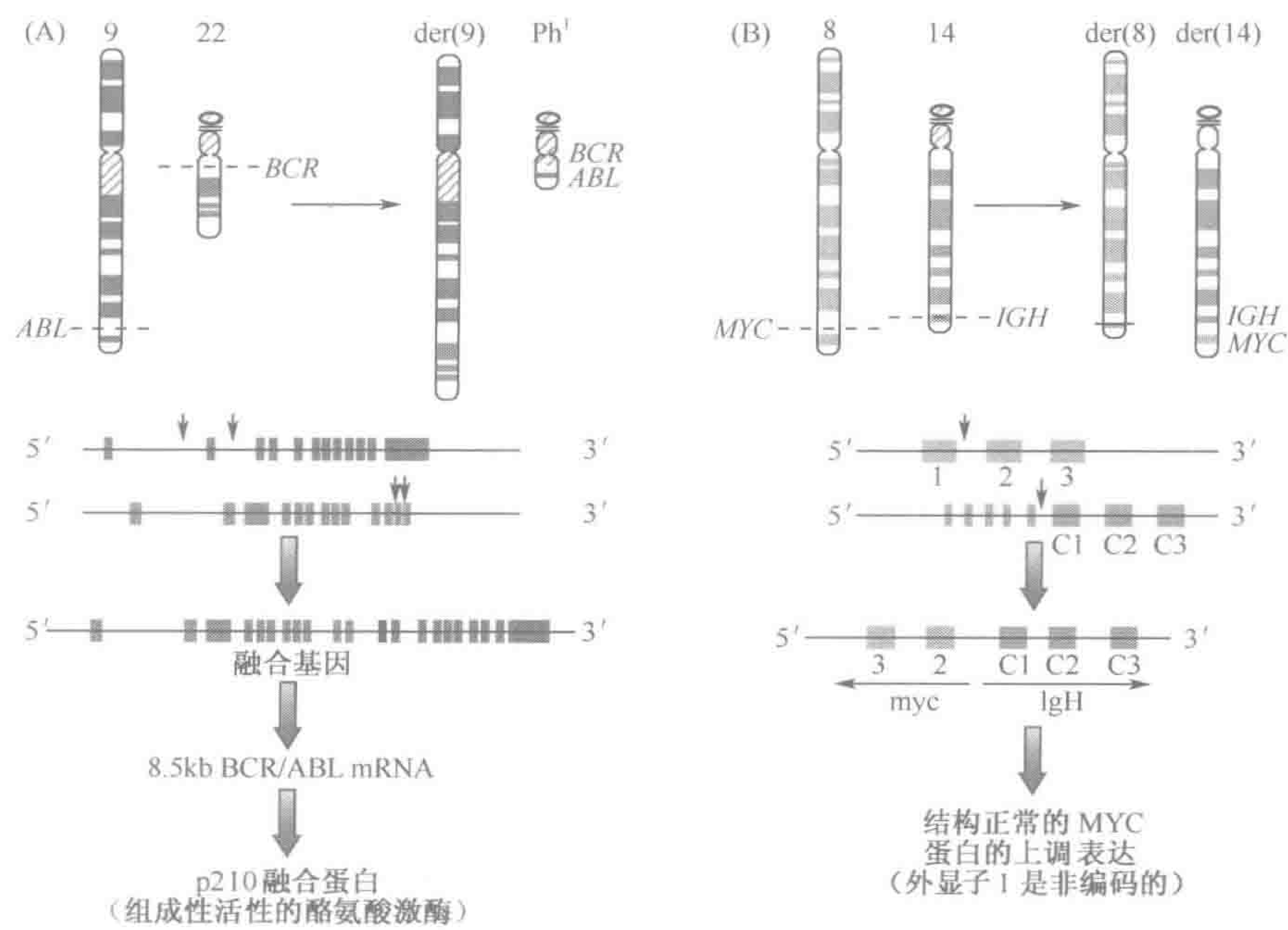


图 17.4　激活癌基因的染色体重排

（A）由慢性髓性白血病 t(9;22) 易位造成的质变所激活。费城染色体上的嵌合型 *BCR-ABL* 融合基因编码一种对正常调控无应答的酪氨酸激酶。详见 Chissoe 等（1995）。（B）由 Burkitt 淋巴瘤 t(8;14)易位造成的量变所激活。8 号染色体上的 *MYC* 基因易位至免疫球蛋白重链基因内。在 B 细胞中该区域的转录活跃，引起 *MYC* 的过度表达。

由易位至转录活跃的染色质区激活

Burkitt 淋巴瘤是常见于非洲中部和巴布亚新几内亚岛疟疾区的一种儿童肿瘤。蚊子和 EB 病毒被确认为在该病的病因学中起一定的作用，然而，*MYC* 癌基因的激活是

其一个核心事件。一种特征性的染色体易位 t（8;14)(q24;q32）可见于该病 75%～85%的患者中（图 17.4B)。其他患者则具有 t（2;8)(p12;q24）或者 t（8;22)(q24;q11）易位。其中，每一种易位都将 *MYC* 癌基因置于某一免疫球蛋白位点附近，如 14q32 上的 *IGH*、2p12 上的 *IGK* 或 22q11 上的 *IGL*。与表 17.3 所示的肿瘤特异染色体易位不同，Burkitt 淋巴瘤易位不产生新的嵌合基因。相反，它们把癌基因置于在产生抗体的 B 细胞中某一活跃转录的染色质环境中。通常，*MYC* 基因第一外显子（为非编码序列）未参与易位。由于缺少自身正常的上游控制并被置于某一活跃的染色质区中，因此，*MYC* 基因处于一种不恰当的高表达水平。然而，这也许并非是事情的全部。这些易位产生于涉及免疫球蛋白 V-D-J 基因重排（节 10.6）的特殊重组酶的误导作用，而且，易位的 *MYC* 基因经常包括新的诱导点突变作为产生抗体多态性的部分机制。

由相同机制产生的许多其他染色体重排将某一癌基因置于某免疫球蛋白（*IGG*）基因或 T 细胞受体（*TCR*）附近（Sanchez-Garcia，1997)。可以预见，这些重排是白血病和淋巴瘤，而不是实体瘤的特征。

表 17.3　由肿瘤特异性染色体重排所产生的嵌合基因

肿瘤	重排	嵌合基因	嵌合型产物的性质
CML	t(9;22)(q34;q11)	*BCR-ABL*	酪氨酸激酶
尤文氏肉瘤	t(11;22)(q24;q12)	*EWS-FLI1*	转录因子
尤文氏肉瘤(变异型)	t(21;22)(q22;q12)	*EWS-ERG*	转录因子
软组织恶性黑色素瘤	t(12;22)(q13;q12)	*EWS-ATF1*	转录因子
促结缔组织生成性小圆细胞瘤	t(11;22)(p13;q12)	*EWS-WT1*	转录因子
脂肉瘤	t(12;16)(q13;p11)	*FUS-CHOP*	转录因子
AML	t(16;21)(p11;q22)	*FUS-ERG*	转录因子
乳头状甲状腺癌	Inv(1)(q21;q31)	*NTRK1-TPM3*(*TRK* 癌基因)	酪氨酸激酶
B 细胞前体急性淋巴细胞性白血病	t(1;19)(q23; p13.3)	*E2A-PBX1*	转录因子
ALL	t(X;11)(q13;q23)	*MLL-AFX1*	转录因子
ALL	t(4;11)(q21;q23)	*MLL-AF4*	转录因子
ALL	t(9;11)(q21;q23)	*MLL-AF9*	转录因子
ALL	t(11;19)(q23;p13)	*MLL-ENL*	转录因子
急性早幼粒细胞性白血病	t(15;17)(q22;q12)	*PML-RARA*	转录因子+维甲酸受体
肺泡横纹肌肉瘤	t(2;13)(q35;q14)	*PAX3-EKHR*	转录因子

CML：慢性髓性白血病；ALL：急性淋巴细胞性白血病；AML：急性髓性白血病。
注意，相同的基因可能涉及几种不同的重排。详见 Rabbitts (1994)。

17.4　肿瘤抑制基因

17.4.1　视网膜母细胞瘤范例

我们所了解的肿瘤抑制基因背景是由 Stanbridge（1990；见进一步阅读）所描述的。一个早期的里程碑是 Knudson 于 1971 年对视网膜母细胞瘤（Knudson 综述，2001）的研究。这是一种由一群分裂快速和分化差的过渡性细胞团所引发的攻击性儿童癌症，正如我们较早使用的公共汽车比喻，这些细胞团已极具危险，因此，转化相对容易。大约 40%的病例是家族性的。它们以不完全外显的显性方式遗传（MIM 180200）。家族性病例通常是双侧性的，而散发病例总是单侧性的。Knudson 指出双侧性病例的发病年龄分布与单一突变一致，而散发病例遵循二次打击规律。他推论到，所有视网膜母细胞瘤涉及二次“打击”，但在家族性病例中一次打击是遗传性的（图 17.5）。Cavenee 等（1983）对生殖的研究既证明了 Knudson 假说，又树立了对肿瘤抑制基因实验研究的范例。

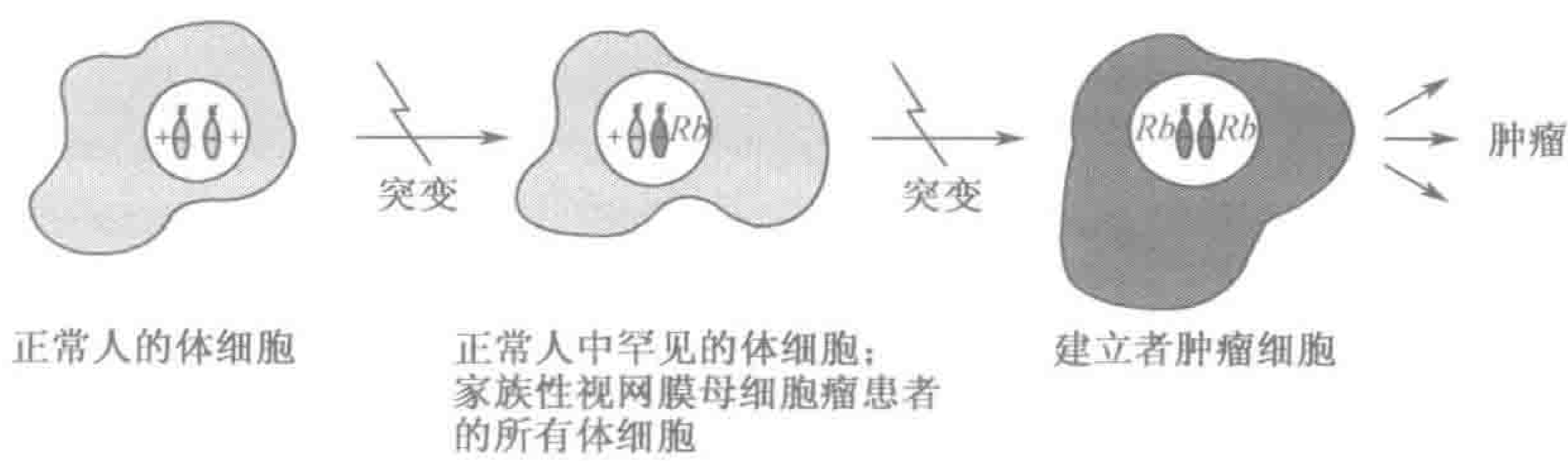


图 17.5　Knudson 的两次打击假说

假设有 100 万个靶细胞，每个细胞的突变概率为 10^{-5}。散发型视网膜母细胞瘤需要两次打击，将影响万分之一的个体（$10^6\times10^{-5}\times10^{-5}=10^{-4}$），而家族性视网膜母细胞瘤仅需要一次打击，外显率相当高，由于 $10^6\times10^{-5}>1$。详情可参见 Knudson（2001）。

Cavenee 及其同事应用一系列 13 号染色体（已知视网膜母细胞瘤有时与染色体 13q14 的变化相关）标记，对散发性视网膜母细胞瘤患者手术切除的瘤组织进行了分类研究。当比较同一患者外周血和肿瘤标本的分析结果时，他们注意到，在几个病历中，对某一或多个标记来讲，结构（外周血）DNA 为杂合性，而肿瘤细胞 DNA 很明显为纯合性。推测他们所见到的是其中一次 Knudson“打击”：肿瘤抑制基因中的一个功能等位基因丢失。随后的研究证实了这种解释，因为在遗传性病例中，总是野生型等位基因以这种方式丢失。将 13q 不同区域的标记研究与细胞遗传学分析结合起来，Cavenee 等提出了缺失发生的几种机制（图 17.6）。

这项工作有两个重要涵义，首先认为，散发性和家族性癌可能拥有共同的分子机制。其次，提供了两种发现肿瘤抑制基因的途径。

- 家族性肿瘤相关基因的作图和定位克隆；
- 肿瘤特异性染色体物质的扫描。

表 17.4 列举了通过罕见家族性癌研究所鉴定的一些肿瘤抑制基因。

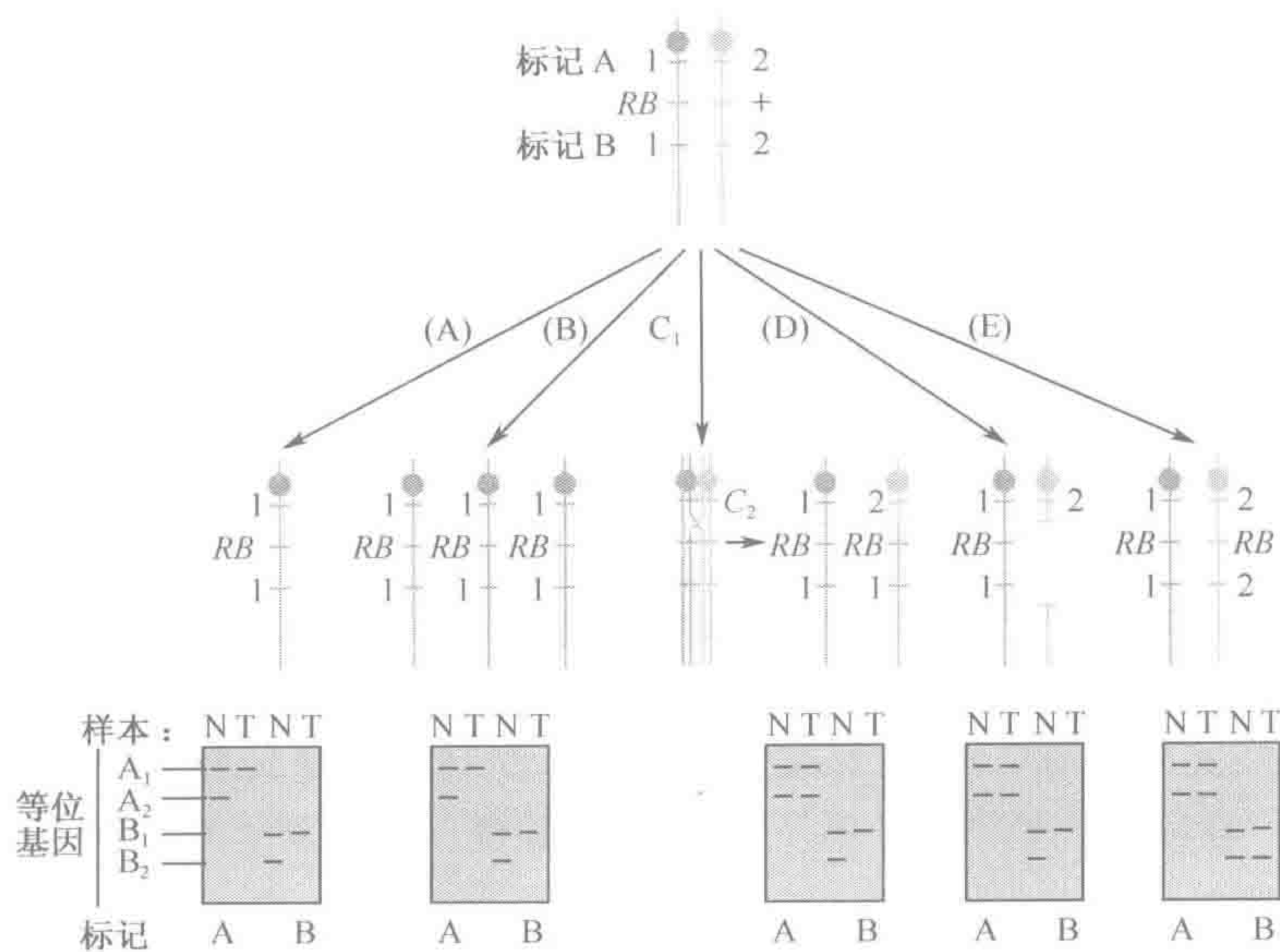


图 17.6　视网膜母细胞瘤中野生型等位基因丢失的机制

(A) 由有丝分裂不分离引起的整条染色体的丢失。(B) 丢失之后染色体复制（在本例中）产生三个拷贝的 Rb 染色体。(C) 临近 Rb 基因座的有丝分裂重组（C_1），之后两个拥有 Rb 的染色体进入同一个子细胞（C_2）；这是对人类或甚至对哺乳动物中的有丝分裂重组的首次证实。(D) 野生型等位基因的缺失。(E) 野生型等位基因的致病性点突变（源自 Cavenee *et al.*，1983)。图下部显示两个标记 A 和 B 的正常（N）与肿瘤（T）DNA 基因型。注意杂合性丢失的类型。

表 17.4　由肿瘤抑制基因突变引起的罕见家族性癌

疾病	MIM 号	染色体定位	基因
家族性腺瘤样结肠息肉	175100	5q21	*APC*
遗传性非息肉性结肠癌	120435,12036	2p16,3p21.3	*MSH2*,*MLH1*
乳腺-卵巢癌	113705	17q21	*BRCA1*
乳腺癌(早期)	600185	13q12-q13	*BRCA2*
Li-Fraumeni 综合征	151623	17p13	*TP53*
Gorlin's 基底细胞痣综合征	109400	9q22-q31	*PTC*
共济失调性毛细血管扩张	208900	11q22-q23	*ATM*
视网膜母细胞瘤	180200	13q14	*RB*
纤维神经瘤 1 型(von Recklinghausen 病)	162200	17q21-q22	*NF1*
纤维神经瘤 2 型(前庭神经鞘瘤)	101000	22q12.2	*NF2*
家族性黑色素瘤	600160	9p21	*CDKN2A*
von Hippel-Lindau 病	193300	3p25-p26	*VHL*

通过 OMIM 索引号可以查到这些基因和疾病的参考文献。Futreal 等（2001）的表 1 给出了一个更长的名单。

视网膜母细胞瘤范例的复杂性

回顾过去，视网膜母细胞瘤是一个非同寻常的、明确的二次打击机制的例子，正如前述，可能因为视网膜母细胞瘤的细胞是不寻常的。许多其他肿瘤的研究的确严格遵循视网膜母细胞瘤的这个范例。如 *APC*、*NF2* 和 *PTC*（表 17.4）就是通过对罕见家族性癌的研究鉴定的，而且，证明在相应的散发性肿瘤中也具有重要作用——但事情并非总是如此。

- 在家族性和散发性病例中，某些肿瘤似乎遵循不同的发生途径。例如，*BRCA1* 仅在 10%～15%的散发型乳腺癌中失活，那些肿瘤则形成乳腺癌中一个可识别的、独特的乳腺癌分子集合。当失活确实存在时，它并非由于在视网膜母细胞瘤（图 17.6）中所见到的染色体机制，而是由于 DNA 的甲基化（节 17.4.3）。
- 肿瘤中某些基因经常丢失其中一个等位基因的功能，而其余等位基因似乎具有完整的功能。有时，第二个等位基因由于 DNA 甲基化而失活，而后者无法被一些实验方法检测到，因此，某些病例似乎仅发生真正的一次打击事件。并非没有理由推测，存在单倍剂量不足（节 16.4.2）但足以提供生长优势的基因。
- 更非寻常的是某些病例存在三次打击。在家族性腺瘤样息肉病（节 17.5.3）中，某些生殖细胞突变程度“弱”，形成了削弱的、无息肉的表型。源于这样患者的腺瘤除遗传性突变外，通常还存在两个体细胞的突变。一个体细胞突变使野生型等位基因失活。这赋予了细胞生长优势，然而，由于染色体缺失造成部分功能性生殖细胞 *APC* 等位基因的丧失，它们获得了进一步的优势。

17.4.2 杂合性丢失检测广泛应用于肿瘤抑制基因的定位

Cavenee 等（1983）发现了视网膜母细胞瘤中体细胞遗传改变是如何引起 *RB* 基因座附近标记的杂合性丢失的。尽管染色体（畸变）机制可能不同，但是对其他肿瘤的研究证实了这一大致情况（Thiagalingam *et al*.，2001）。因此，通过使用跨越基因组的标记来筛查配对的血液和肿瘤样本（通常为散发肿瘤），我们希望能发现肿瘤抑制基因所处的座位（图 17.7）。

在过去的十年里，杂合性丢失分析已经成为肿瘤研究的主要手段，但是，人们不禁要问其结果是否已经证明了这种努力的合理性。对有意义的结果，必须用相距很近的标记来筛查大量的肿瘤。使用具有高度多态性的微卫星标记以尽量减少不提供信息病例的数量，在这些病例中结构 DNA 的标记为纯合子。或者也可以使用微阵列-比较基因组杂交技术（图 17.3B）。晚期癌细胞可能表现出多达 1/4 基因座的杂合性丢失，因此，需要使用大量肿瘤样本从总的背景中排除特殊的变化。这些研究结果提示存在非常大量的肿瘤抑制基因，图 17.8 显示了一个典型的例子。然而，试图遵循这种定位克隆研究的努力，成功率相当低。

图 17.9 举例说明了分析杂合性丢失数据时出现的一个陷阱。一个更普遍的问题是，肿瘤典型地具有数不清的数目和结构的一般异常核型（图 17.10）。几乎得不到任何有关用于杂合性丢失研究的肿瘤核型的信息，但是，图 17.8 所提供的结果容易产生一个自然而然的假设，即特殊的间隙型缺失影响了另一个正常的染色体。事实上，杂合性丢

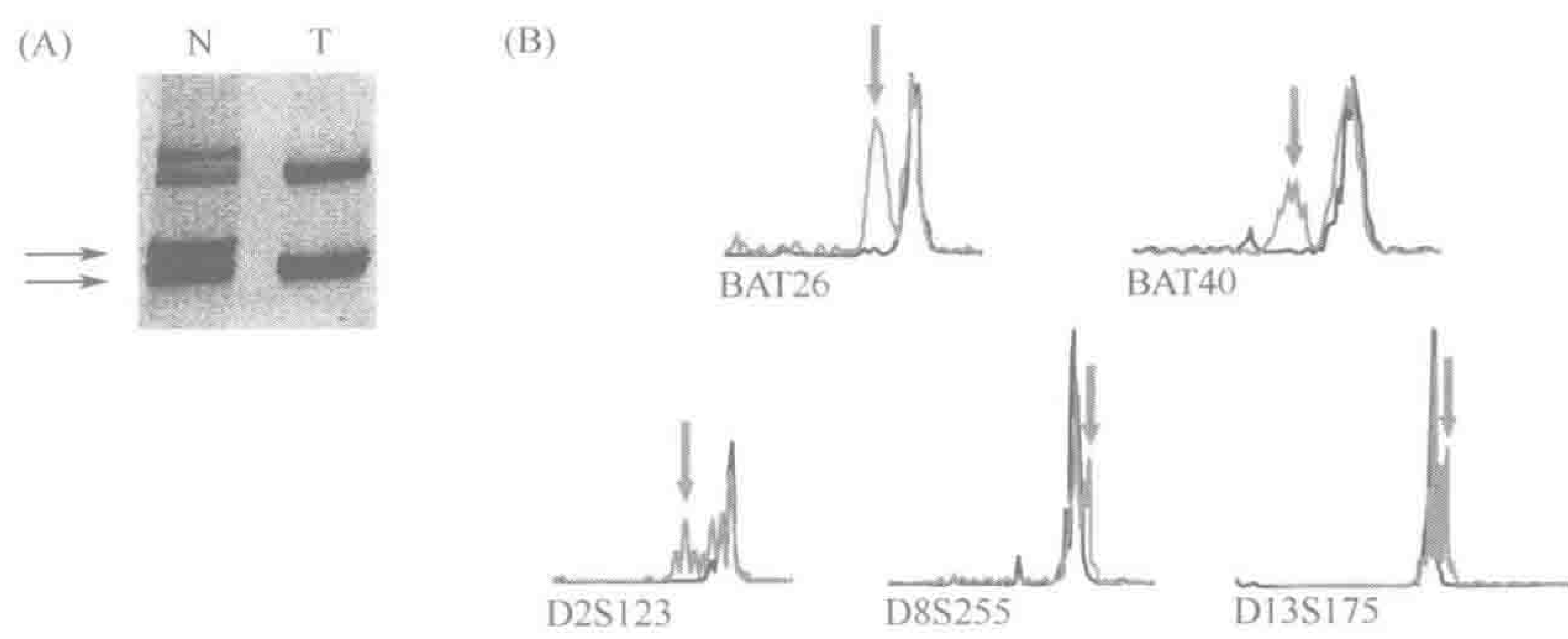


图 17.7　肿瘤的遗传学变化

(A) **杂合性丢失** (loss of heterozygosity)。正常组织标本 (N) 的 *D8S522* (箭头) 标记是杂合性的，而肿瘤标本 (T) 已经丢失了上面的等位基因。更上面的带是"构像带"，是由每个等位基因的折叠序列所产生的辅助带。图片由 Nalin Thakker 博士惠赠，St Marys Hospital，Manchester，VK。(B) 遗传性非息肉性结肠癌中的**微卫星不稳定性** (microsatellite instability)。对遗传性非息肉性结肠癌中的五个微卫星多态进行荧光测序分析。黑色踪迹来源于组成型 DNA，红色踪迹来源于相同患者的肿瘤 DNA。箭头示肿瘤中新的等位基因。*BAT26* 和 *BAT40* 为单腺苷酸 $(A)_n$ 重复多态标记，*D2S123*、*D8S255* 和 *D13S175* 为双核苷酸 $(CA)_n$ 重复。由 Yvonne Wallis 惠赠，West Midlands Regional Genetics Laboratong，Birmingham，VK。

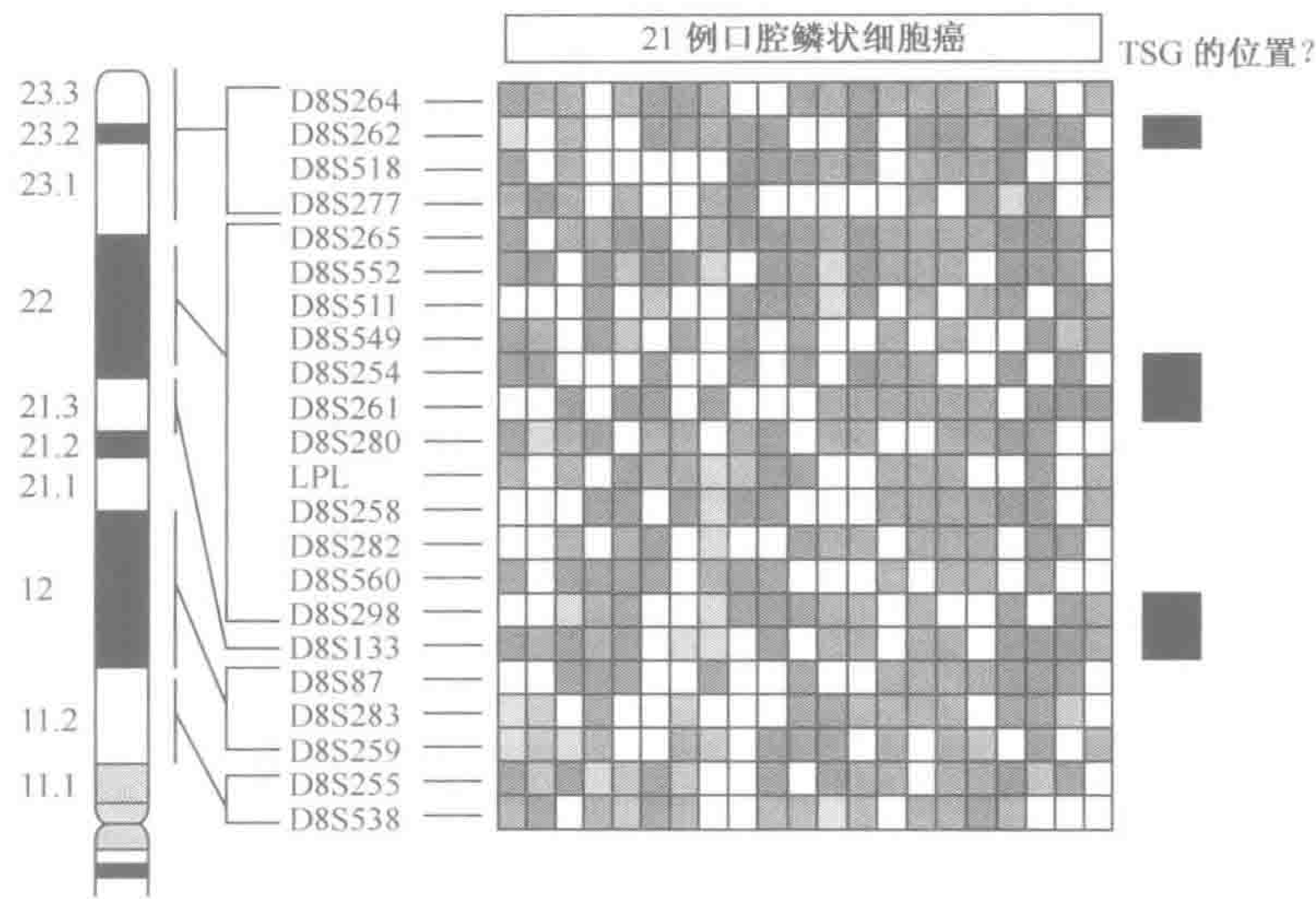


图 17.8　染色体 8p 上可能的肿瘤抑制基因

图中的网格显示应用染色体 8p 上的多种标记 (*D8S264* 等) 对一系列口腔肿瘤患者的结构和肿瘤 DNA 进行分型的结果。粉红色表示杂合性丢失，绿色表示正常，黄色表示结果不明确，灰色表示微卫星不稳定性，由于结构的纯合性，空白处表示在此处标记不提供信息。注意在这些肿瘤中杂合性丢失的复杂模式，后者是此类研究中的典型代表。这些数据提示，在 8p 上存在 3 个不同的肿瘤抑制基因。

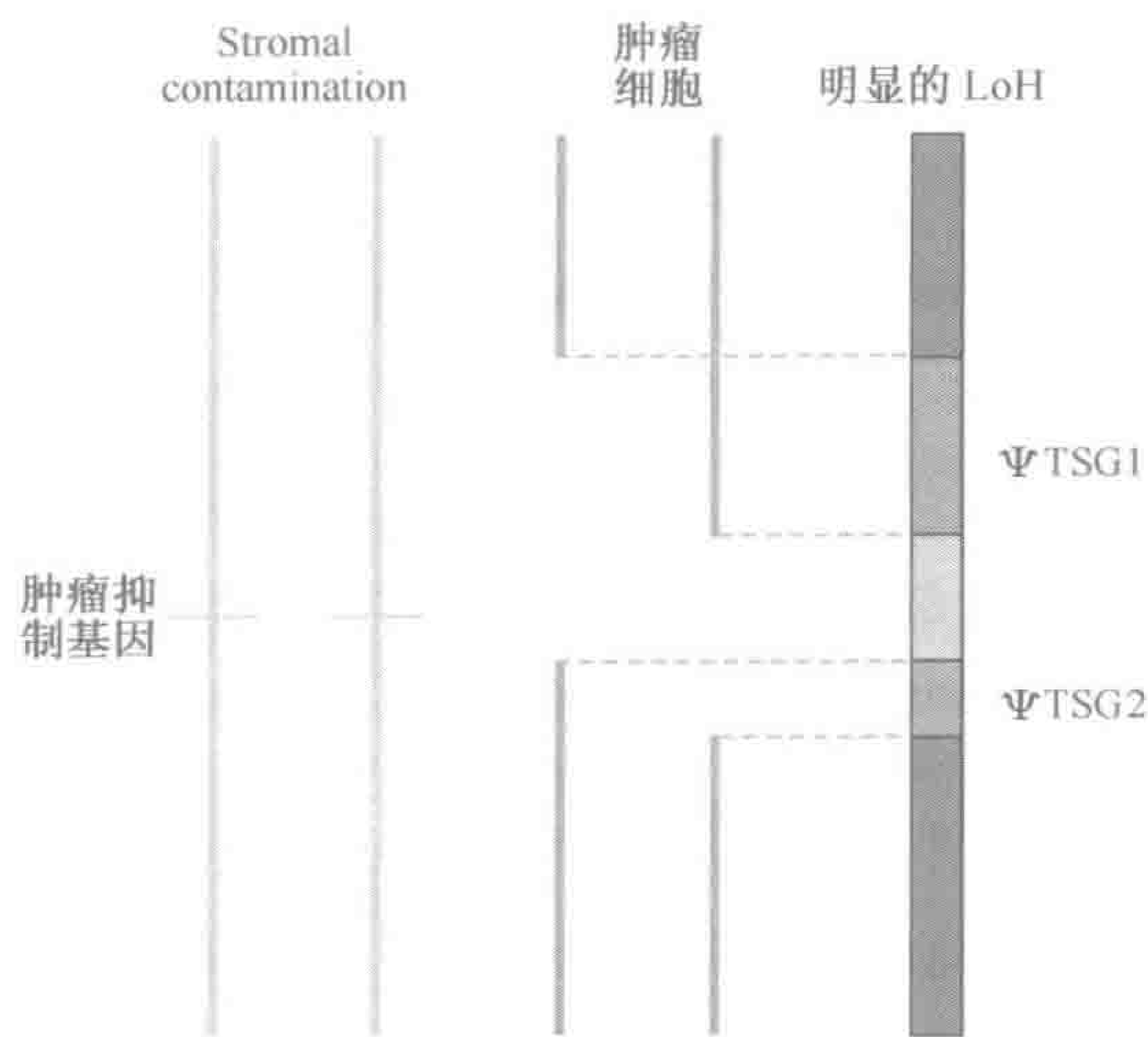


图 17.9 分析杂合性丢失数据时的一个陷阱

该肿瘤中的真实情况是肿瘤抑制基因的纯合性丢失。由于间质的扩增（浅色线），杂合性丢失数据显示肿瘤抑制基因座上的杂合性保留，而在两个侧翼区存在杂合性丢失。这种情况推测出存在两个假肿瘤抑制基因座，ΨTSG1 和 ΨTSG2。真实情况可以由荧光原位杂交或肿瘤抑制基因产物的免疫组织化学检测所揭示。

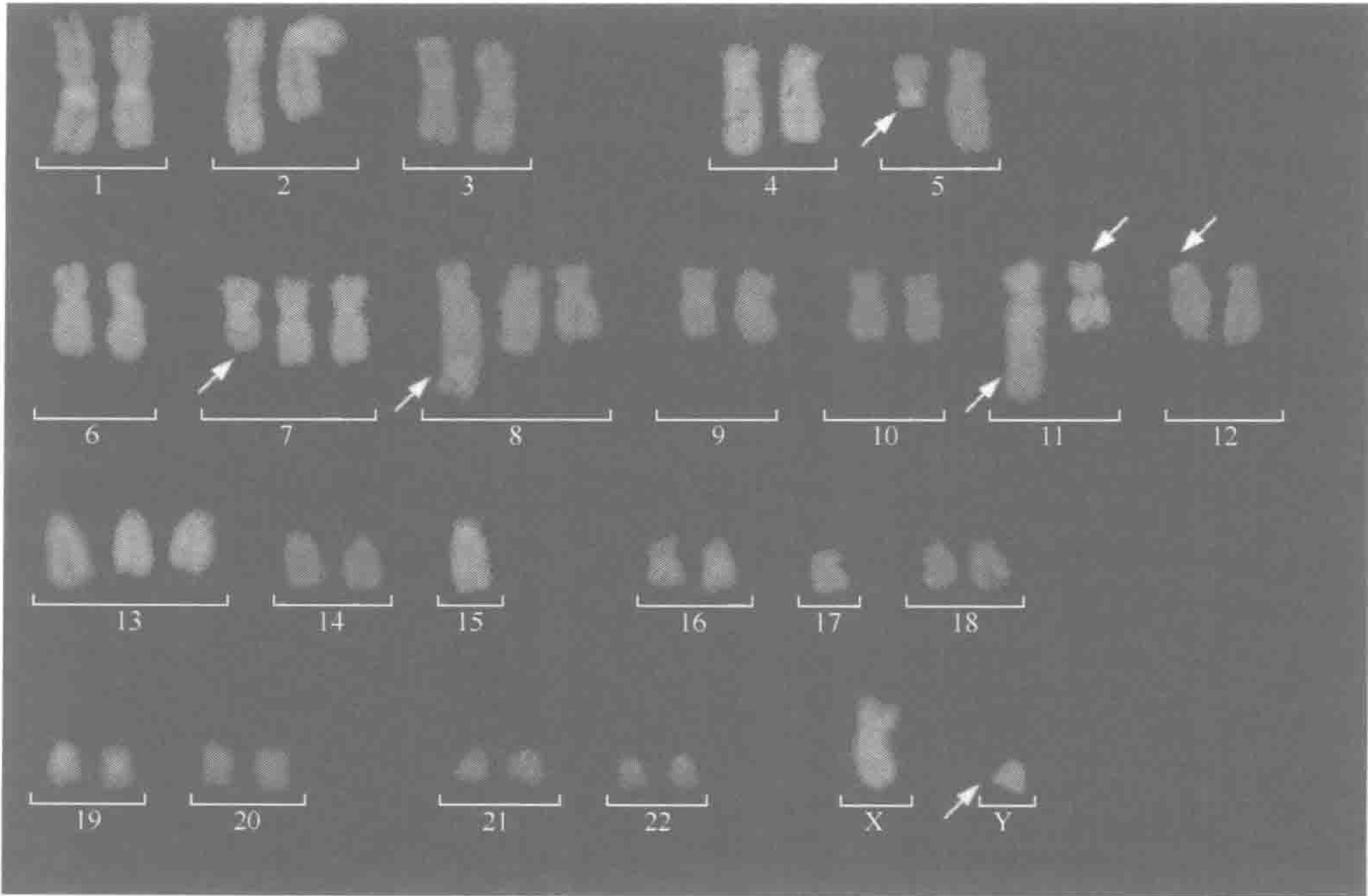


图 17.10 一个人类髓性白血病衍生细胞系的多色 FISH（multicolor FISH，M-FISH）核型

注意由 24 种颜色的染色体涂染所显示的大量数目和结构异常。它们包括 t（5;15）、der（7）t（7;15）、der（8）ins（8;11）、der（11）t（8;11）、t（11;17）和 der（Y）t（Y;12）——后者在这个细胞中不甚清晰。图片由英国牛津分子医学研究所 Lyndal Kearney 博士惠赠，自 Tosi 等，经 John Wiley & Sons，Inc. 的子公司——Wiley-Liss Inc. 的同意而重印。

失是染色体不稳定和具有肿瘤细胞特征性双链 DNA 断裂缺陷修复的产物。所观察到的一些丢失也许是特殊类型染色体不稳定性的副产物，并非是对肿瘤抑制基因丢失的选择。

17.4.3 肿瘤抑制基因常因甲基化而发生表观遗传性沉默

肿瘤抑制基因可能因丢失（表现在杂合性丢失）或点突变而被沉默，但是，第三个极为常见的机制是启动子区甲基化（节 10.4）。整体上，与正常细胞 DNA 相比，肿瘤细胞 DNA 是低甲基化的，但是，基因启动子区特异性 CpG 二核苷酸的甲基化几乎见于所有人类肿瘤，而且与基因不恰当的转录沉默有关。Jones 和 Baylin（2002）列举了很多例子。对一些肿瘤抑制基因而言，甲基化经常发生于点突变，而在其他情况（如 3p21 上的 *RASSF1A*、17p13.3 上的 *HIC1*）下，甲基化是特异肿瘤功能丢失的唯一已知机制。标准突变筛查技术漏检了这些改变，因此，它们的重要性可能还是被低估了。

17.5 基因组的稳定性

基因组不稳定性几乎是肿瘤细胞的一个普遍特征。不稳定性可以有两种类型：

- **染色体不稳定性**（chromosomal instability，CIN）是最常见的类型。肿瘤细胞通常具有一般的异常核型（图 17.10），包括多种额外的和丢失的染色体及很多重排等；
- **微卫星不稳定性**（microsatellite instability，MIN）是一类 DNA 水平的不稳定性，见于少数肿瘤，特别是结肠癌（图 17.7B）。

不稳定性可能是使细胞积聚足够突变所必需的（框 17.1）。有些人争论道，不稳定性只是肿瘤发生中一个伴随的副产物，而且，即使具有标准的突变率，上皮细胞分裂的次数也足以满足所需的累计的突变次数。然而，正常情况下肿瘤或者表现出 CIN，或者表现出 MIN，但不能同时出现这两种事件，由此推断，不稳定性不是一个随机的特征，而是选择的结果。

17.5.1 染色体的不稳定性

见于肿瘤细胞中的许多染色体畸变多数是随机性的，尽管它们为进一步选择生长旺盛的细胞提供了基础。它们可能产生于以下三种途径。

- 肿瘤细胞丢失**纺锤体检查点**（spindle checkpoint）（Jallepalli and Lengauer，2001；见下文）。这可能是很多染色体数目异常的主要来源。
- 尽管存在 DNA 损伤，肿瘤细胞也能经历细胞周期。染色体结构异常可能是含有损伤 DNA 的 DNA 复制和有丝分裂的副产物。
- 有时，肿瘤也可能复制达到使端粒变短至不能保护染色体末端的水平，这将产生各种各样的结构异常（见下文）。

纺锤体检查点

纺锤体检查点应该阻止染色体在有丝分裂中的分离，直至所有染色体均正确地附着在纺锤丝上。其分子机制并不十分清楚。已经鉴定了几个候选基因，但仅一个在癌细胞

中经常突变的基因是 *APC*，该基因涉及结肠息肉病。*APC* 基因编码一个大的多功能蛋白质，后者可能参与各种细胞过程。在结肠，CIN 甚至见于非常早期的腺瘤，而且 $APC^{-/-}$ 细胞存在异常的有丝分裂纺锤体，从而引起染色体不稳定。

DNA 损伤信号系统

细胞不断地修复各种 DNA 损伤。对这种损伤的正常反应就是终止细胞周期直至损伤被修复，而许多癌的一个普遍特征是丧失了这种调控。这种令人困惑的、复杂的检测系统和信号 DNA 损伤从人到酵母大多是保守的（Zhou and Elledge，2000）。这种调控涉及到一个称作 BRCA1 相关基因组监视复合体（BRCA1-associated genome surveillance complex，BASC）的多蛋白装置与大量的其他蛋白。几个已知的肿瘤抑制基因出现在这个装置中：

- ATM 是 BASC 的成分之一，也是损伤敏感机制的一个早期成分。这种巨大蛋白将信号传递给其下游的靶蛋白。ATM 功能的丧失引起共济失调性毛细血管扩张症（ataxia telangiectasia，AT；MIM 208900）。受累纯合子易患各种的肿瘤、免疫缺陷症、染色体不稳定和小脑共济失调。某种特异 ATM 突变的杂合子携带者患乳腺癌的风险增加；
- nibrin 与 MRE11 和 RAD50 蛋白构成复合体。这种复合体参与形成 BASC 的部分结构，也许还有其他功能。缺乏 nibrin 则会引起 Nijmegen 断裂综合征（Nijmegen breakage syndrome，NBS，MIM 251260）。临床上，NBS 与 AT 非常相似，但是，它还包括巨头畸形和生长迟缓而无共济失调。*NBS* 基因的克隆见节 13.5.2 中的描述；
- BRCA1，家族性乳腺癌（节 15.6.1）中发现的第一个基因产物，是又一个非常大的具有多功能结构域的蛋白质，参与形成 BASC。BRCA1 还具有重组、染色质重塑和转录调控等功能（Scully and Livingston，2000）。
- BRCA2 在结构上与 BRCA1 并不相似，但是，却拥有许多相同的功能。某种特殊类型的 *BRCA2* 突变既是一些遗传性乳腺癌的病因，又可以引起 B/D 型 Fanconi 贫血（MIM 227650）。这是一种常染色体隐性综合征，包括先天性畸形、进行性骨髓衰竭、细胞对 DNA 损伤的高敏感性以及对肿瘤的易感性。

这些蛋白缺陷的细胞修复双链断裂的能力似乎特别差。其他类型的损伤修复很大程度上可能不依赖这个系统的检测。

端粒和染色体的不稳定性

正如我们在节 2.2.5 中所见，人类染色体的末端受重复序列（TTAGGG）n 的保护，这个重复序列由一个特殊的含 RNA 的酶系统——端粒酶维持。端粒酶存在于人类的生殖细胞中，而大多数体细胞组织缺乏这种酶，端粒的长度在每代细胞中缩短 50～100bp。经长时间培养后，正常细胞趋于衰老，这时它们将停止分裂。p53、视网膜母细胞瘤蛋白 pRb 缺陷的或含有病毒癌基因的成纤维母细胞不衰老而持续分裂，并出现危机。大多数细胞死亡，但是，存活下来的千万分之一左右的细胞拥有一般的（gross）染色体畸变，获得了端粒酶并发生永生化。可能有这样一种“危机”，即端粒太短以至于不能再保护染色体末端而避免重排。

肿瘤细胞中一般染色体畸变的一种机制可能由于过度的细胞分裂而使端粒的消耗达到危机的程度（Maser and DePinho，2002）。然而，这并不能解释持续存在的不稳定性，因为肿瘤细胞总是要以某种形式获得维持其端粒并无限复制的能力。85%～90%的成熟转移癌具有不断更新的端粒酶活性；其余的则具有另一种不同的称为 ALT 并以重组为基础的机制。

17.5.2 DNA 修复缺陷和 DNA 水平的不稳定性

如上所示，细胞在不断地修复其 DNA 损伤（Hoeijmakers，2001），而且，修复系统的缺陷是一系列易感癌症的遗传性疾病产生的基础。主要缺陷如下：

- 核苷酸切除修复缺陷——由电离辐射、紫外线或化学致癌剂损伤造成的 DNA 单链断裂和交叉在进入下一轮复制前需要被修复（节 11.6.1）。某些修复酶的缺陷见于几种易感肿瘤综合征，特别是各种类型的着色性干皮病（XP；MIM 278700）。XP 患者是在遗传上有功能突变缺失的纯合体，无法修复由紫外线引起的 DNA 损伤。他们对日光过度敏感，并在暴露的皮肤上发生许多肿瘤；
- 碱基切除修复缺陷——缺陷的碱基切除修复在人类癌症中很少被人注意，但是，某种类型的结肠癌是由 MYH 修复酶的缺陷所引起的（Al-Tassan *et al.*，2002）；
- 双链断裂修复缺陷-双链断裂由同源重组或非同源末端连接来修复（Hoeijmakers，2001）的。这两种方法都需要上述 ATM-NBS-BRCA1-BRCA2 机制；
- 复制错误修复缺陷——这些缺陷是在结肠癌的研究中被发现的，如下所述。

17.5.3 遗传性非息肉性结肠癌和微卫星不稳定性

多数结肠癌是散发的。家族性病例分为两类。

- **家族性腺瘤样息肉病**（FAP 或 APC；MIM 175100）是一种常染色体显性遗传病，成百上千个息肉遍布于患者的结肠。息肉（腺瘤）不是恶性的，但是，如果不治疗的话，其中的一个或多个息肉必然会发展成有侵袭性的癌。肿瘤抑制基因 *APC* 的遗传性突变是该病的病因（表 17.4）。
- **遗传性非息肉性结肠癌**（HNPCC；MIM 120435，120436）也呈常染色体显性遗传，外显率非常高，然而，与 FAP 不同，它没有息肉阶段。HNPCC 基因定位于两个位点，2p15-p22 和 3p21.3。

关于 HNPCC 的杂合性丢失研究发现了新的、意想不到的结果。肿瘤样本并非缺乏结构上的 DNA 等位基因，而是表现出新增的微卫星标记等位基因。图 17.7B 显示了这样的例子。杂合性丢失是特异染色体区的一个特性，但是，微卫星不稳定（MIN）是 HNPCC 的普遍现象。许多肿瘤偶然地表现出一个或几个微卫星多态的不稳定性（见图 17.8 中的例子），但是，高频率的不稳定性（简单地定义为受检的所有标记应该大于 29%；Tomlinson *et al.*，2002）定义了一类具有不同临床病理学特征的 MIN^+ 肿瘤。

在一个横向思维的范例中，Fishel 等（1993）叙述了大肠杆菌和酵母中被称作增变基因的 MIN^+ 现象。这些基因编码一种纠正错误的系统，该系统检查新合成 DNA 的错配碱基对或小的插入-缺失环。编码大肠杆菌系统的 *MutHLS* 基因的突变导致突变率增加 100～1000 倍。Fishel 及其同事克隆了一个与这些基因同源的人类基因 *MutS*，发现

该基因位于 2p 上一个 HNPCC 基因的座位，并在一些 HNPCC 家族中表现出结构突变。总的来说，与大肠杆菌基因同源的 6 个基因与人类的错配修复有关（Jiricny and Nystrom-Lahti，2000）；见表 17.5 和图 17.11。

表 17.5 参与 DNA 复制错误修复的基因

大肠杆菌	人类	人类染色体定位	HNPCC 中所占百分比
MutS	*MSH2*	2p16	35%
	MSH3	5q11－q12	0%?
	MSH6	2p16	5%[a]
MutL	*MLH1*	2p21.3	60%
	MLH3	14q24.3	0%?
	PMS2	7p22	很低

a 非典型的晚发型 HNPCC 和子宫内膜癌。详见 Jiricny 和 Nystrom-Lahti（2000）。

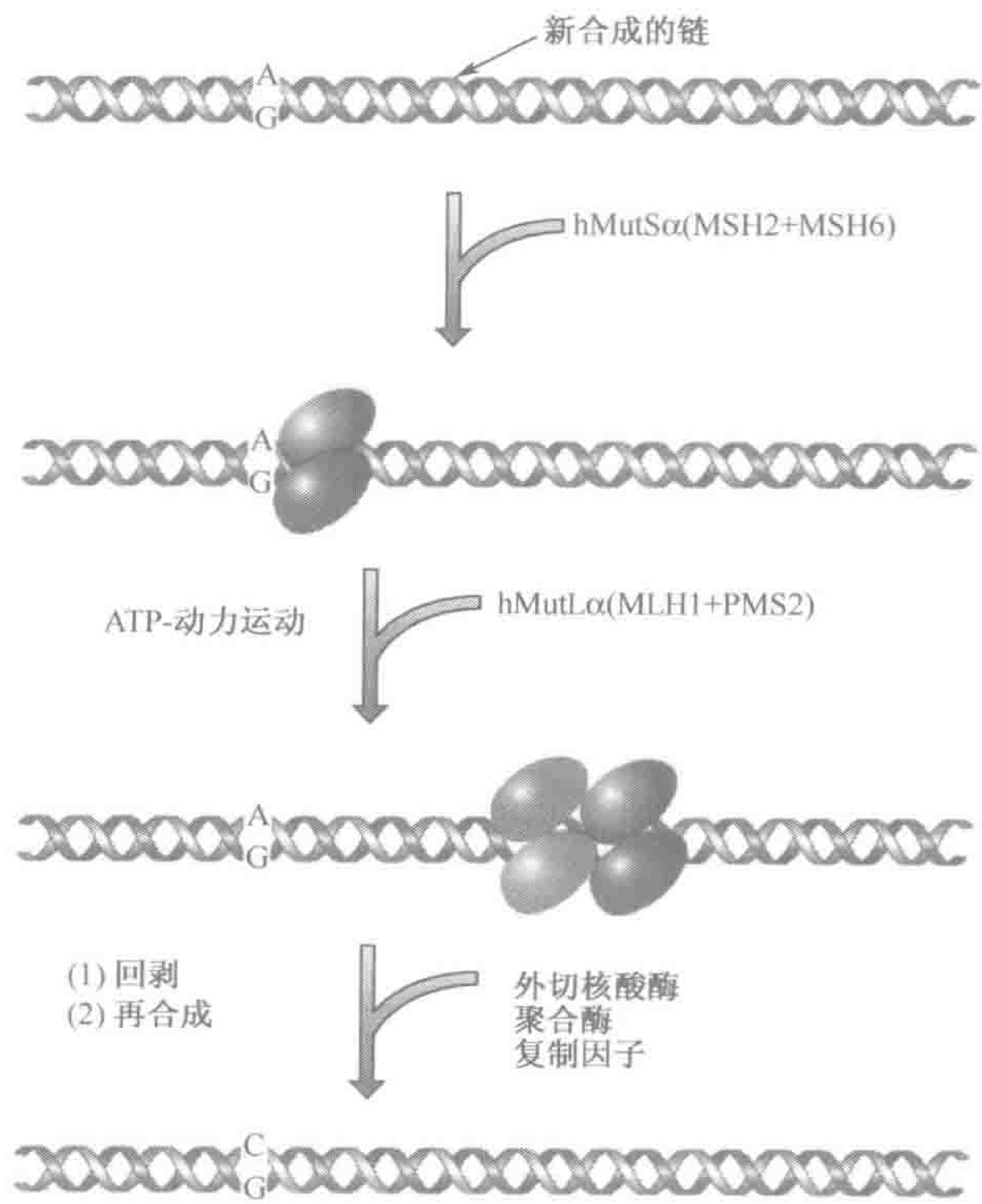


图 17.11 错配修复机制

复制错误可以导致错配的碱基对或小的插入/丢失环。它们被 MSH2/MSH6 二聚体 hMutSα 或有时为 MSH2/MSH3 二聚体 hMutSβ 所识别。这些蛋白沿 DNA 易位，并结合 MLH1/PMS2 二聚体 hMutLα，然后组装成完整的“修复体”，回剥错配的碱基，重新合成新链。见 Jiricny 和 Nystrom-Lahti（2000）。

HNPCC 患者几乎总是 *MLH1* 和 *MSH2* 基因功能丢失突变的组成型杂合子。他们的正常细胞仍然具有功能性的错配修复系统，但是，不表现出 MIN^+ 表型。在肿瘤中，第二个等位基因的丢失是由如图 17.6 所示的机制之一引起的。微卫星不稳定性见于 10%～15%的结肠癌、子宫内膜癌和乳腺癌中，但是，在其他肿瘤中仅偶见发生。

17.5.4 p53 和细胞凋亡

基因组不稳定的主要原因是 *TP53* 基因的丢失或突变，该基因编码转录因子 p53。这种丢失可能是癌中最常见的单一遗传变化。这反映在 p53（称其为基因组卫士）的核心作用（Vousden，2000）。如上所述，细胞周期停止于 DNA 损伤的细胞。如果损伤无法修复，将触发细胞凋亡。p53 在此过程中发挥关键作用。正常情况下，细胞中 p53 处于低水平，因为该蛋白被快速降解。来自所有的细胞压力感受器包括损伤感受器的信号引起 p53 的磷酸化和稳定化。这增加了依赖 p53 的基因转录，如抑制细胞周期的 $p21^{WAF1/CIP1}$ 基因和控制细胞凋亡的 *PUMA*、*PIG3* 基因。缺乏 p53 的肿瘤细胞可以继续复制损伤的 DNA，而且不经历细胞凋亡。不过，要注意 p53 丢失是肿瘤发生相对晚期的事件（见图 17.17 中的例子）；证据表明肿瘤发生的早期不受 p53 保护。

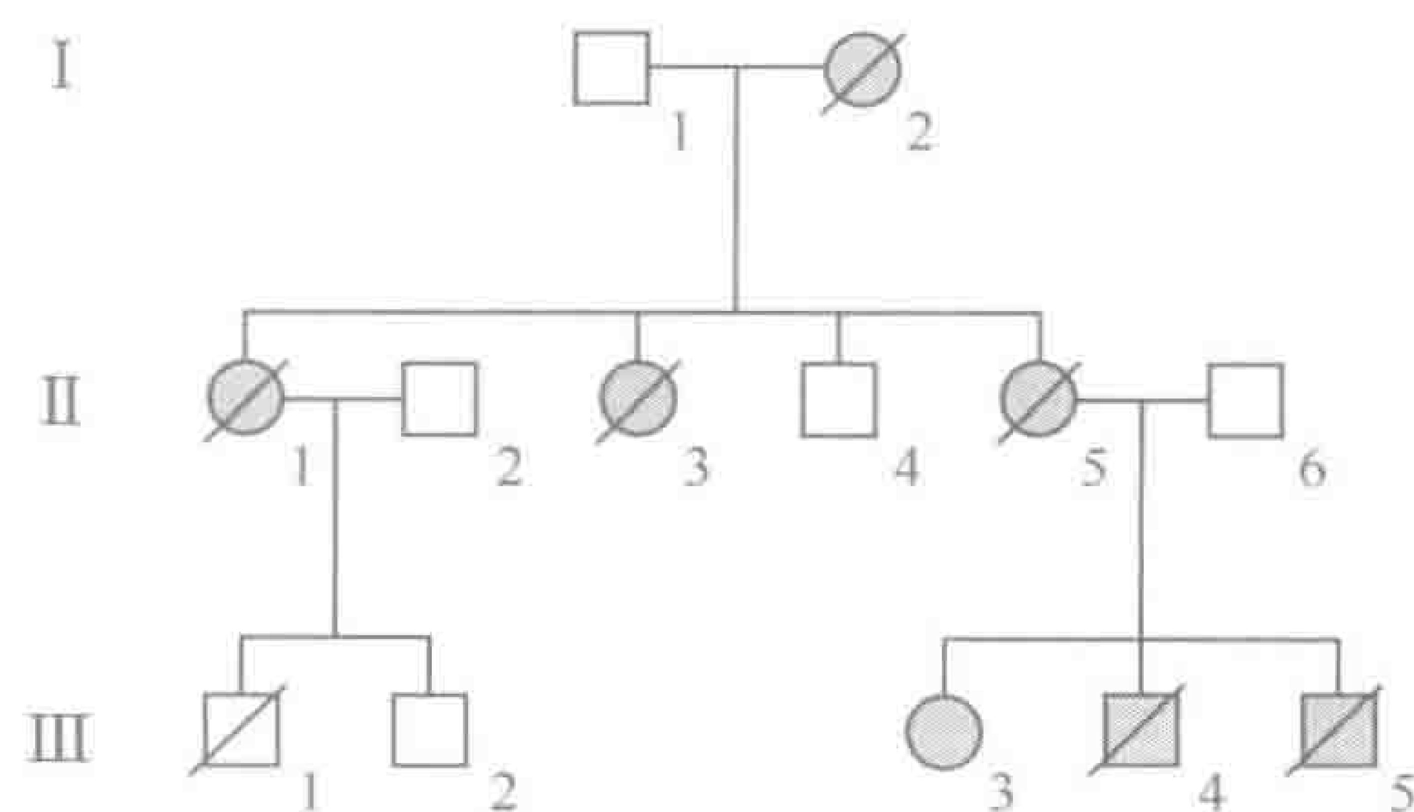


图 17.12 Li-Fraumeni 综合征的一个典型系谱

Li-Fraumeni 综合征典型的恶性表现包括，40 岁时被诊断为双侧乳腺癌（Ⅰ-2）；35 岁时被诊断为脑肿瘤（Ⅱ-1）；19 岁时被诊断为软组织肉瘤和 33 岁时被诊断为乳腺癌（Ⅱ-3）；32 岁时被诊断为乳腺癌（Ⅱ-5）；8 岁时被诊断为骨肉瘤（Ⅲ-3）；2 岁时被诊断为白血病（Ⅲ-4）；3 岁时被诊断为软组织肉瘤（Ⅲ-5）。Ⅰ-1 在 59 岁时被诊断为结肠癌——经推测，该病与 Li-Fraumeni 综合征无关。系谱源自 Malkin（1994）。

p53 可以被丢失、突变或如 *MDM2* 基因产物（该产物与 p53 结合并使之成为降解的靶标；该产物还与 pRB 结合，见下文）或乳头状病毒 E6 蛋白的抑制子的作用所敲除。*TP53* 位于 17p12，这是很多肿瘤发生杂合性丢失最常见的区域之一。无 *TP53* 丢失的肿瘤经常为突变类型。为了弄清作为肿瘤抑制基因的 *TP53*，该基因的结构性突变在显性遗传的 Li-Fraumeni 综合征（MIM 151623）家族中被发现。受累家族成员患多种原发性肿瘤，典型的有软组织肉瘤、骨肉瘤、乳腺肿瘤、脑和肾上腺皮质肿瘤以及白血病（图 17.12）。

17.6 细胞周期的调控

在任何时候，每个细胞有 3 种行为选择：它可以保持静止、分裂或死亡（凋亡）状态。一些细胞还可以选择分化状态。细胞选择其中一种行为以响应内部和外部的信号（图 17.13A）。癌基因和肿瘤抑制基因在产生和解释这些信号中起主要作用。

如果信号与应答之间由单线性的通路（图 17.13B）相连的话，生命就简单多了，然而似乎从未如此。相反，众多分支、重叠以及部分丰余的通路控制细胞的行为（图

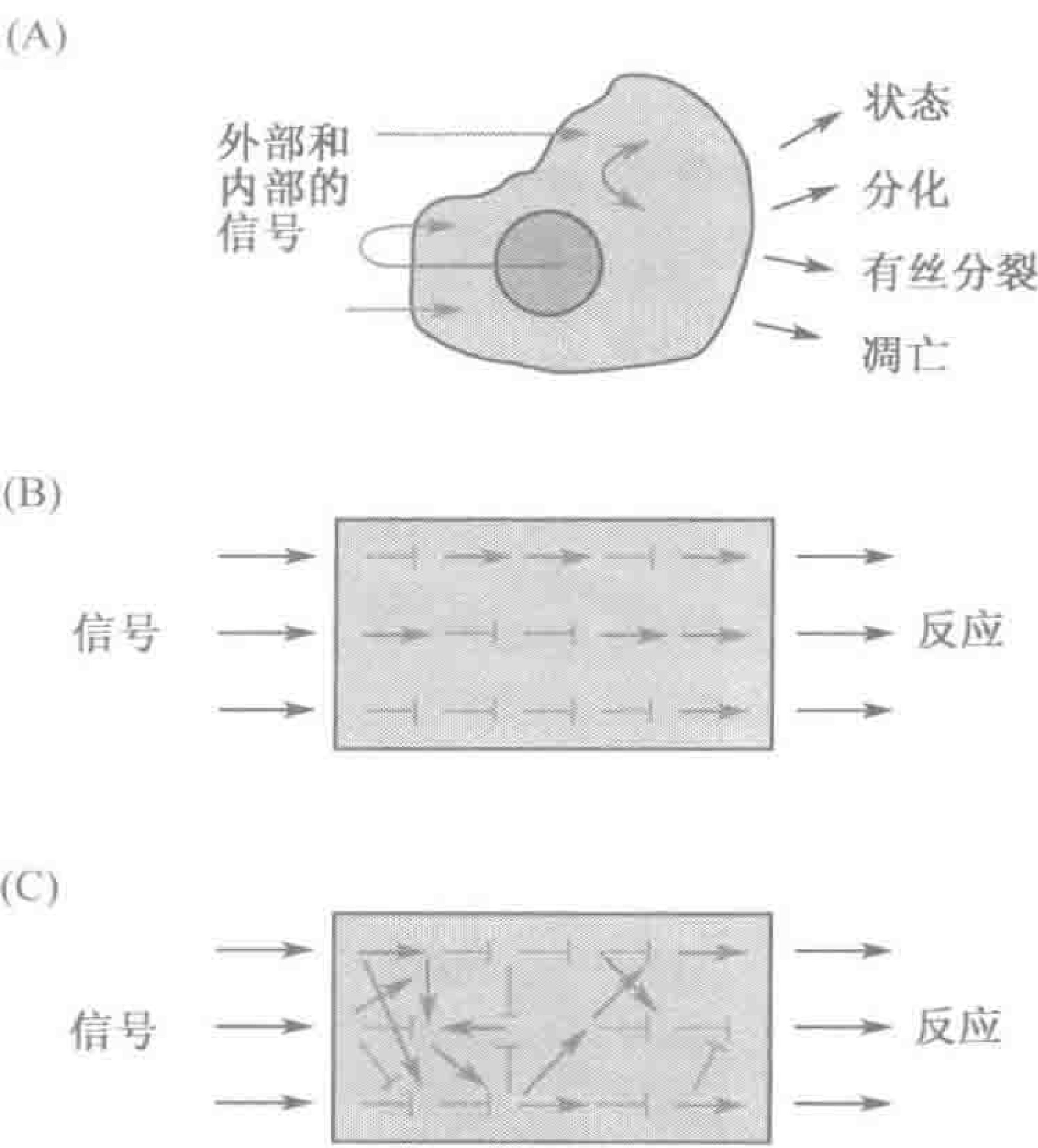


图 17.13 一个细胞面临选择时将如何处之

(A) 细胞在静止、有丝分裂、细胞凋亡和有时为分化状态之间进行选择，以回应内部和外部的信号。(B) 在一个假想的细胞中，信号由线性、非分支的刺激（→）或抑制（⊣）通路连接起来进行应答反应。人类细胞不像这样工作。(C) 在真实的细胞中，信号进入一个部分过度相互作用的复杂网络，其结果不易由分析来预测。

17.13C)。这种复杂的网络可能赋予细胞独特复杂机制的稳定性和恢复能力。从实验角度来看，弄清精确的遗传调控通路相当难，部分是由于它们的复杂性，部分是由于难以区分转染或敲除实验中的直接与间接效应。希望表达微阵列将成为解决此问题的钥匙。

对于一些细胞如选择分裂的癌细胞来讲，经历细胞周期的过程易受几个检查点的影响（图 17.14）。三个主要的检查点如下：

- G_1-S 检查点——当存在未修复的 DNA 损伤时，DNA 复制被阻止；无法修复的损伤导致细胞凋亡。在 S 期内可能有一个额外的独立损伤检查点；
- G_2-M 检查点——细胞被阻止进入有丝分裂期直至完成 DNA 复制和损伤修复；
- 纺锤体检查点——见节 17.5.1 中的描述。

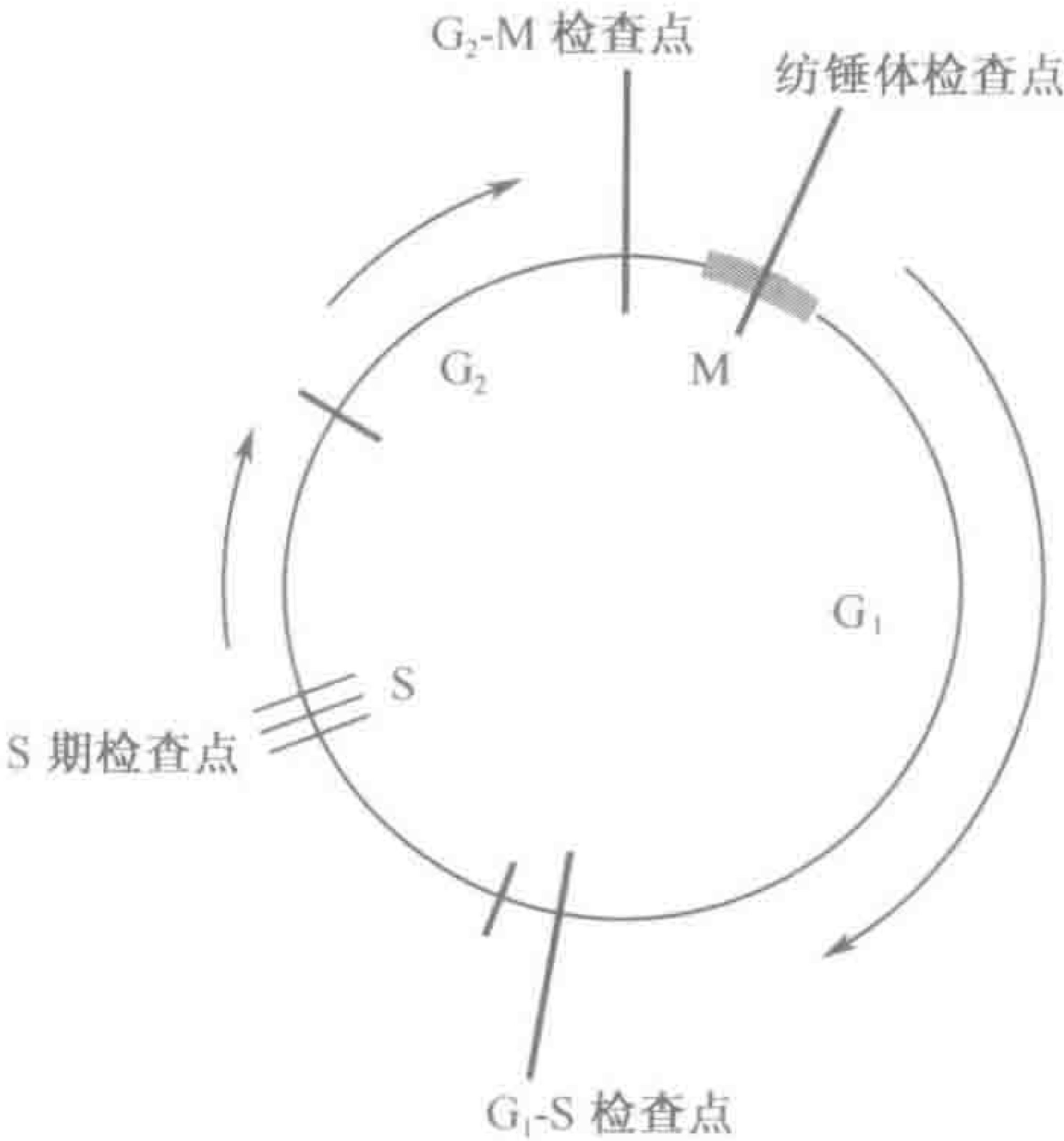


图 17.14 细胞周期检查点

一系列的检查点阻止细胞随未修复的 DNA 损伤进入 S 期或有丝分裂期，或阻止细胞进入有丝分裂前期直至所有染色体都正确地与纺锤丝相连。在 S 期，可能存在一个额外的 DNA 损伤检查点。

17.6.1　G1-S 检查点

图 17.15 显示该检查点的部分路径。此检查点似乎非常重要，因为在多数肿瘤中该检查点是失活的。三个关键的肿瘤抑制基因 *RB*、*TP53* 和 *CDKN2A* 在肿瘤发生中处于核心地位，而且是肿瘤细胞中最常见的出现异常的基因。每个基因在遗传性癌中也发挥作用。事实上，可能所有的肿瘤细胞需要同时使该系统的 RB 和 p53 的分支失活，以便细胞绕过细胞周期的常规检查，并避免触发细胞凋亡以回应未修复 DNA 损伤或过度生长的信号。

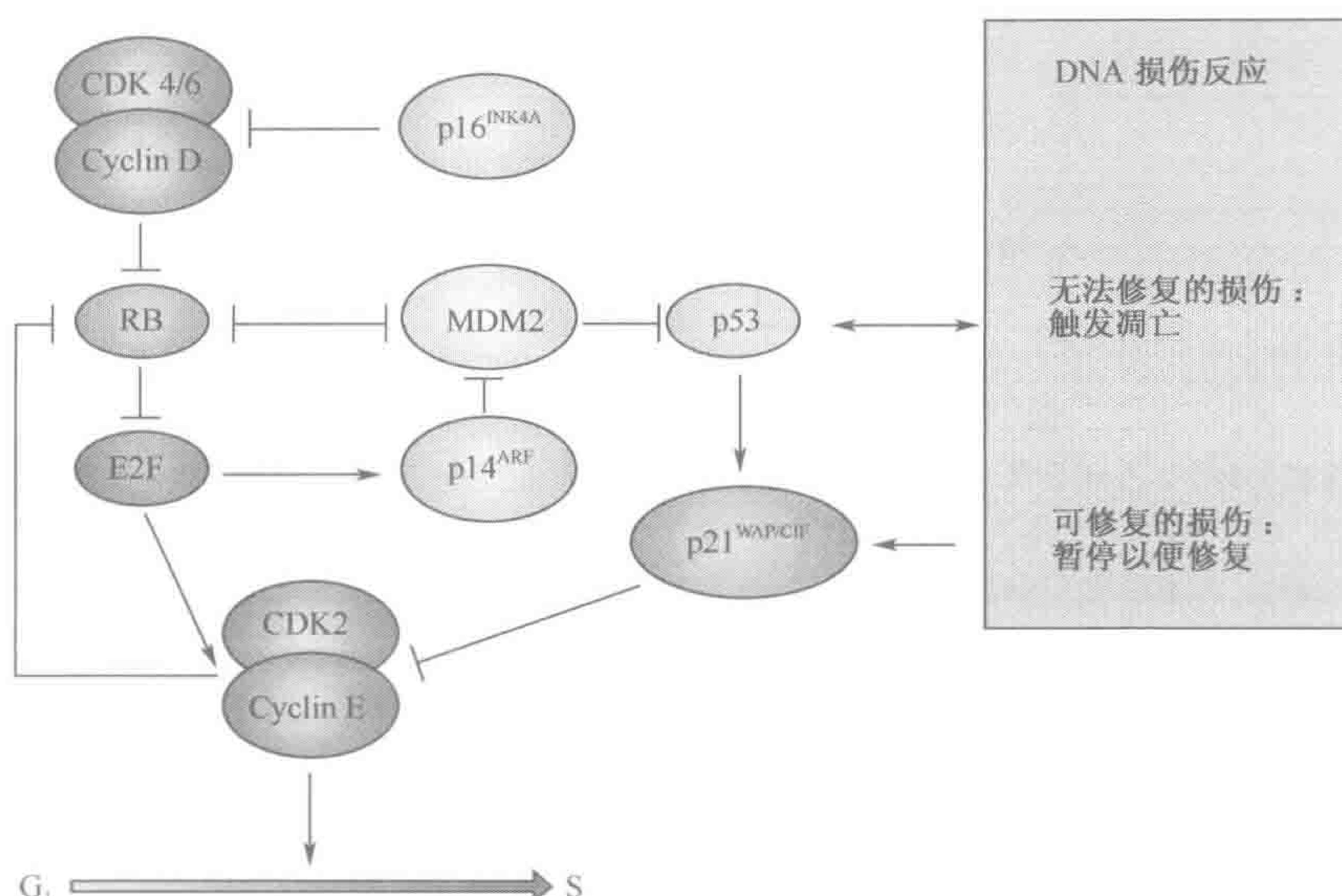


图 17.15　由 *RB*、*TP53*、*CDKN2A*（INK4A）基因产物介导的对细胞周期进展和基因组完整性的控制

这些控制至少构成部分 G1-S 细胞周期检查点。详见 Harbour 和 Dean（2000）。

RB 基因产物——pRb 的功能

RB 基因是通过其在视网膜母细胞瘤中的作用而被鉴定的，但是，它表达广泛，因此有助于对所有细胞的细胞周期的控制。在正常细胞中，该基因产物为一个 110kDa 的核蛋白，能被磷酸化失活并被去磷酸化激活。活化（去磷酸化）的 pRb 结合细胞转录因子 E2F 并使之失活，后者的功能是细胞周期进展所必需的（图 17.15；详见 Harbour and Dean，2000）。在细胞进入 S 期前 2～4 小时，pRb 被磷酸化。这解除了对 E2F 的抑制，使细胞进入 S 期。磷酸化受控于一个由细胞周期素、细胞周期素依赖性激酶和细胞周期素激酶抑制子所组成的级联。几种病毒癌蛋白（腺病毒 E1A、SV40-T 抗原、人类乳头状病毒 E7 蛋白）结合并螯合或降解 pRb，因而促进细胞周期的进展。

CDKN2A 基因产物 p16^{INK4A}和 p14^{ARF}的功能

位于 9p13 的重要基因 *CDKN2A*（以前被称为 *MTS1* 或 *INK4A*）编码两个结构上

不相关的蛋白质（图 17.16）。外显子 1α、2 和 3 编码 INK4A 或 p16。第二启动子在外显子 1β 上游启动转录。外显子 1β 被剪接至外显子 2 和 3，但读框发生了移位，因而编码完全不相关的蛋白质 p14 或 ARF（alternative reading frame；小鼠的同源蛋白质为 p19）。

这两个基因的产物在细胞周期调控中起作用（图 17.15）。p16 在 RB 蛋白的上游发挥作用以控制 G_1-S 检查点。细胞周期素依赖性激酶通过磷酸化使 pRb 失活，而 p16 抑制细胞周期素依赖性激酶 CDK4/6。因此，p16 功能的丢失导致 RB 功能的丢失和不正常的细胞周期。*CDKN2A* 基因的另一个产物 p14，通过使 *MDM2* 癌基因产物 MDM2 蛋白失去稳定性来介导 G_1 期捕获，*MDM2* 癌基因在许多肉瘤（Pomeranz *et al.*，1998）中扩增。MDM2 结合 p53 并诱导其降解。因此，p14 起维持 p53 水平的作用。p14 功能的丢失引起 MDM2 表达水平过高、p53 过度的破坏，从而使细胞周期失去控制。

通常仅影响 p16 的 *CDKN2A* 遗传突变见于一些多发性黑色素瘤，但是，体细胞突变频率非常大。该基因的纯合丢失使细胞周期调控中的 RB 和 p53 分支失活，是肿瘤发生中很常见的现象。某些肿瘤存在影响 p16 而不是 p14 的突变（例如 1α 启动子因甲基化而失活）。这些肿瘤还具有 p53 突变的倾向，表明如图 17.15 所示的控制系统中两个蛋白分支同时突变的重要性。

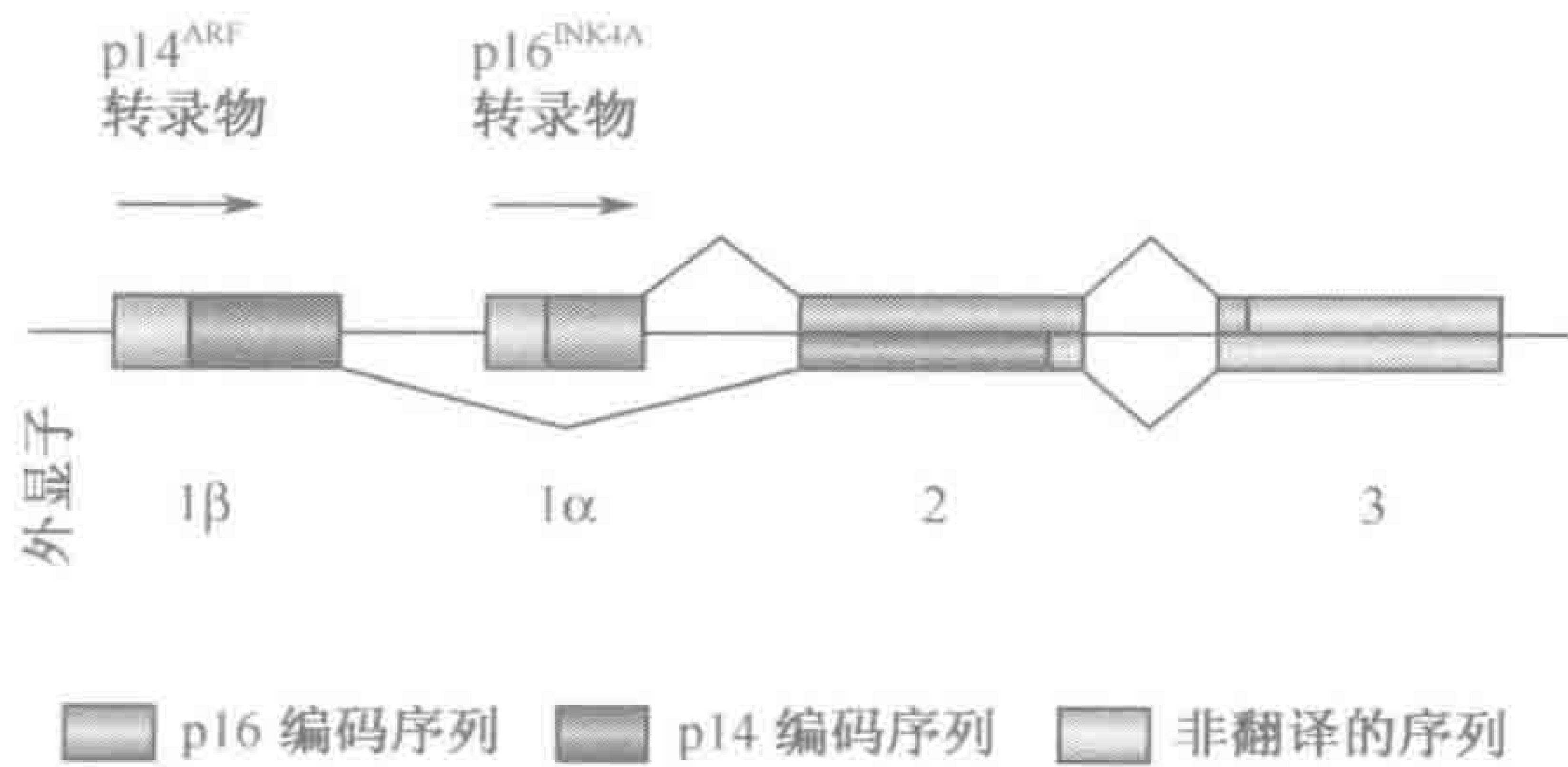


图 17.16 *CDKN2A* 基因的两种产物

该基因（也称为 *MTS* 或 *INK4A*）编码两种毫不相关的蛋白质。$p16^{INK4A}$ 由外显子 1α、2 和 3 编码，$p14^{ARF}$ 由外显子 1β、2 和 3 编码，但是，外显子 2 和 3 具有不同的读框。这两种基因产物分别在细胞周期调控的 RB 和 P53 分支通路中被激活，如图 17.15 所示。

17.7 整合资料：通路和能力

正如我们在本章开头所明确指出的那样，肿瘤遗传学可能特别令人费解。如此多的基因、如此多的突变、如此多的重组……，每个肿瘤都是独特的。但是，它们的共同之处在于每个肿瘤都是对一系列特定的可能性选择的结果，并且，当我们按通路而不是按单个基因来思考问题时，也只有有限的方式来获取这些可能性。Hahn 和 Weinberg（2002）试图列出所有的这些通路并描绘出了癌症的“路线图”。

17.7.1 结肠癌中的通路

我们对肿瘤发生的理解能力的提高源自结肠癌，因为可以应用手术切除的家族性腺瘤样息肉病患者的结肠来研究肿瘤发生的所有阶段。图 17.17 显示了这些阶段。我们可以为其他无丰富数据的肿瘤制定相似但更基本的方案。这样的方案基于一系列的观察：

- 恶性肿瘤是从正常结肠上皮通过微小异常腺管病灶到称为腺瘤的良性上皮增生发展而来的。腺瘤通过早期（小于 1cm）、中期（大于 1cm 但无癌巢）和晚期（大于 1cm 且有癌巢）阶段发展为癌，最终变为转移癌；
- 在 FAP，位于 5q21 上的一个 *APC* 基因发生了结构突变。有趣的是，遗传性突变几乎不能使 APC 蛋白失活；它们通常产生具有显性负作用的截短蛋白（节 16.4.3）。证据表明，*APC* 的一个等位基因简单失活不会提供必需的生长优势。最早检出的病变即异常腺管病灶缺乏 APC 的表达。大约百万分之一的细胞发展成息肉，这个概率与作为决定因素的第二个 *APC* 等位基因缺失的概率一致；
- 大约 50%的中、晚期腺瘤和仅约 10%的早期腺瘤存在 *KRAS* 癌基因（*HRAS* 的同族成员，表 17.1）的突变。因此，*KRAS* 的突变可能通常是早期腺瘤进展到中期腺瘤所必需的；
- 大约 50%晚期腺瘤和癌表现 18q 的杂合性丢失。这种情况在早期和中期腺瘤中相对少见。似乎与之相关的基因是 *SMAD4*（也称作 *MADH4*），而不是最初的候选基因 *DCC*（White，1998）。

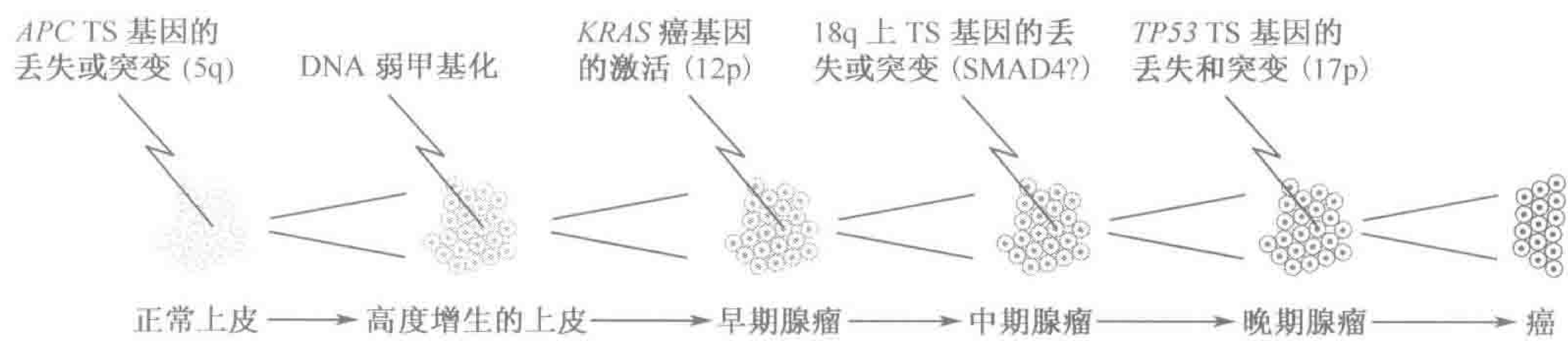


图 17.17　结肠癌的多步骤发生模型

详见正文。这主要是一个思考肿瘤如何发生的方法，并非是一个严谨的描述。每例结肠癌的发生都可能经历相同的组织学阶段，但是，潜在的遗传学变化则差别很大。根据 Fearon 和 Vogelstein（1990）的研究结果，本图显示了一系列特别常见的事件。Smith 等（2002）对该模型的真实性提出了质疑，因为在他们所研究的 106 例肿瘤中，仅 7 例同时发生了 *APC*、*KRAS* 和 *P53* 三个基因的突变。

图 17.17 中的方案试图显示最常见的一系列突变，但是，并非所有结肠癌遵循这一路径。大约只有 60%的结肠癌存在 *APC* 突变。然而，无 *APC* 突变的肿瘤通常存在 β-catenin 的激活突变，后者是 APC 作用的主要下游靶蛋白（Fodde *et al*.，2001）。进一步发展，FAP 肿瘤倾向于丢失 18q，但是，HNPCC 肿瘤都具有稳定的染色体，且无 18q 的丢失。这两种肿瘤的选择压力可能是抑制生长的转化生长因子 β 信号的丧失。在 FAP，这种情况发生于 18q 上的下游效应基因 *SMAD4* 的丢失。在 HNPCC，90%的 MIN^+ 肿瘤在 TGFβ 受体Ⅱ基因内部的 8 个腺嘌呤 $(A)_8$ 中发生了移码突变。HNPCC 肿瘤还经常发生 *AXIN2* 基因的移码突变，该基因编码 APC-β-catenin 通路中的另一种

成分。因此，在分子机制的异质性背后，还存在一个非常规则的通路，该通路必须失活才使肿瘤发生。

17.7.2 一个成功的肿瘤必须获得 6 种特殊能力

一个重要而且被高度推荐的由 Hanahan 和 Weinberg（2000）所著的综述（进一步阅读）提到了另一种思考肿瘤发生的途径。他们指出，从正常细胞到恶性细胞转变中的主要因素就是 6 种特殊能力的获得。细胞必须：

- ▶ 不依赖于外源性生长信号；
- ▶ 对外源性抗生长信号不敏感；
- ▶ 能够逃避细胞凋亡；
- ▶ 具有无限复制的能力；
- ▶ 具有维持血管生成的能力；
- ▶ 具有侵袭组织和转移的能力。

转移可能不是一个特殊的选择能力。对任何细胞而言，转移能力并非是一个明确的优势，也许是晚期肿瘤细胞中普遍的基因组错排的偶然效应。其他 5 种能力则是特殊选择的结果。对有关血管生成的了解相对较少，但是，我们已经看到激活的癌基因、扰乱的细胞周期调控、p53 突变和端粒酶的再激活在获得前 4 种能力中的作用。每种能力的获得需要一系列不同的遗传改变，但是，在每个病例中，获得这种能力需要某一特殊通路的激活或失活。在不同肿瘤中，获得能力的次序不必相同，但是，对中期阶段基因组不稳定性和不断的克隆性扩张（框 17.1）的需要使这个过程呈现某种规律性。

肿瘤多步骤发生中的首次突变至关重要，因为它将赋予其他具有完全防御能力的正常细胞某种生长优势。根据持家基因假说（Kinzler and Vogelstein，1996），在一个既定的、不断更新的细胞群体中，某特异基因承担维持一定细胞数量的任务。某持家基因突变导致细胞分裂和细胞死亡之间的永久不平衡，而如果该基因功能正常，那么其他基因突变将不具备长期效应。按照这个理论，从孟德尔式肿瘤中鉴定的肿瘤抑制基因就是相关组织的持家基因，包括施旺（Schwan）细胞中的 *NF2*、肾细胞中的 *VHL* 等。对 APC 来说，这也许是恰当的，即 APC 蛋白不仅调节细胞增殖的重要控制者 β-catenin 的水平，而且还在纺锤体检查点（节 17.5.1）中发挥重要作用。我们已经注意到，甚至很早期的腺瘤也表现出染色体不稳定性。可能的情况是单个 *APC* 基因突变（合适的类型）赋予结肠上皮某种生长优势，而单个 *BRCA1/2* 基因突变不赋予导管上皮生长优势。这也许能解释为什么 *APC* 基因突变在散发型结肠癌中非常普遍，而 *BRCA1/2* 基因突变在散发型乳腺癌中却非常罕见（Lamlum *et al.*，1999）。像肺癌和前列腺癌这样缺乏孟德尔遗传方式的常见肿瘤如何符合持家基因原理仍不清楚。

17.8 本章所有知识的用途

根据本章提出的思路，显而易见的是现在谈论“肿瘤的治愈”是愚蠢的。然而，最新的知识已经正在以下三个方面取得重要进展：检测、诊断和治疗。

- ▶ 症状出现前癌的检测传统上依赖于体检、扫描（乳房造影法等）、生化方法（前列腺

特异性抗原）或侵入性检测（结肠镜等）。由于不完全明确的原因，肿瘤可脱落DNA。灵敏的PCR方法可以扩增尿液（对于膀胱癌）、粪便（对于结肠癌）、唾液或痰（对于口腔或肺癌）、或血液（对于各种肿瘤）(Sidransky，2002）中的DNA。有希望的是，对*TP53*、*APC*等的突变或特异性启动子甲基化或微卫星不稳定性的DNA检测将成为更有效和无损伤的方法。

▶ 如本章开始所述，肿瘤鉴定和分期的传统组织学方法正不断地被分子的方法所补充（但不是取代）（Lakhani and Ashworth，2001)。这对白血病非常有用，在白血病中，染色体重排的概念为预后和治疗提供了重要的启示（例如 Armstrong *et al*.，

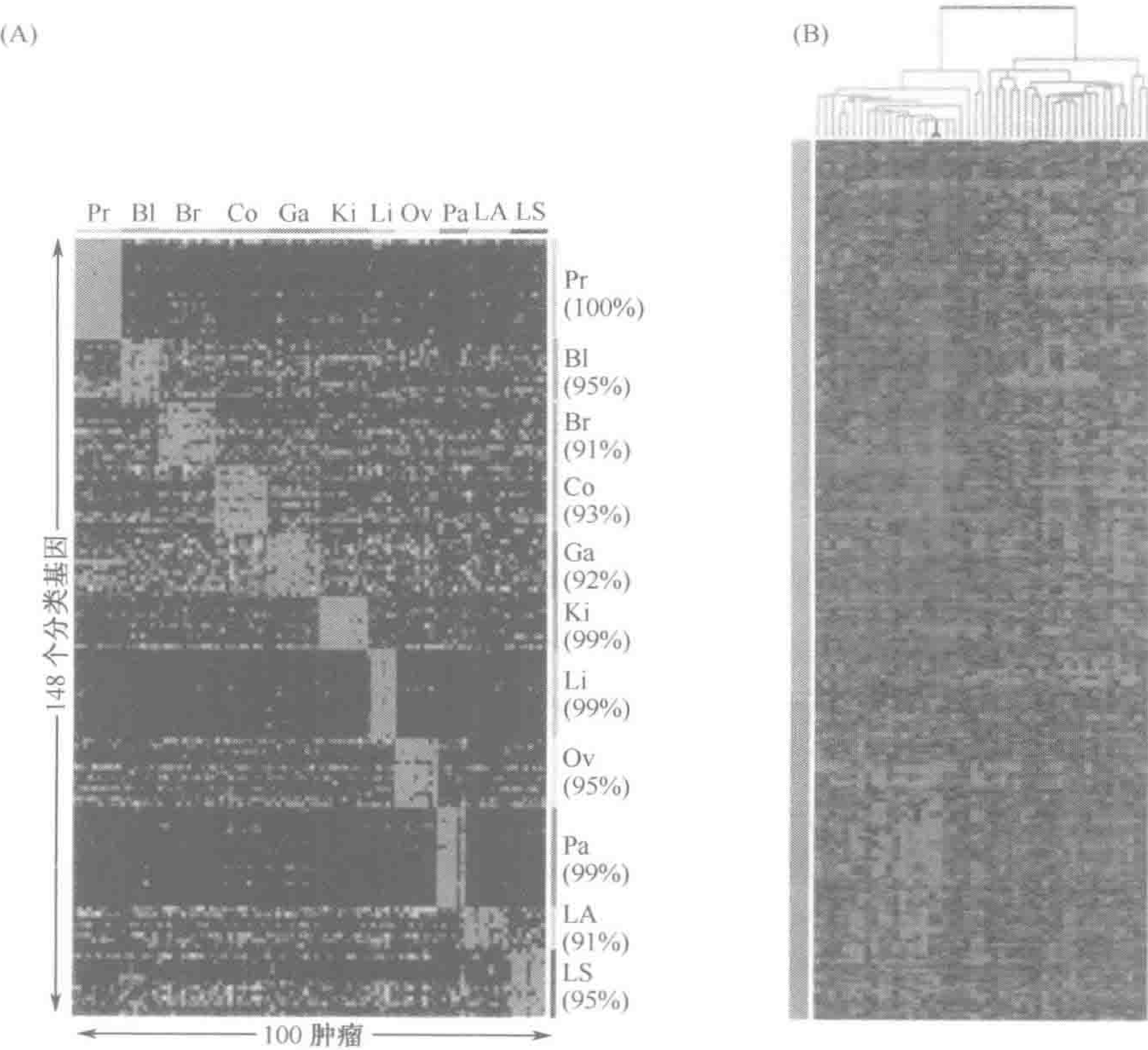


图 17.18 应用表达阵列来刻画肿瘤特征的例子

(A) **组织起源鉴定**。148个基因（行）按来源将100个肿瘤（列）分类的表达谱。Pr：前列腺；Bl：膀胱/输尿管；Br：乳腺；Co：结肠；Ga：胃食管；Ki：肾；Li：肝；Ov：卵巢；Pa：胰腺；LA：肺腺癌；LS：肺鳞癌。红色示基因表达增高，蓝色示基因表达降低。源自 Su 等（2001)，*Cancer Research* 61，7388～7393，经美国癌研究协会允许。(B) **区分两种类型的大的弥散性B细胞淋巴瘤**。基因表达的聚类结果将中心胚样肿瘤激活的B细胞样肿瘤区分开（橘色和蓝色条）。两组的五年存活率分别为76%和16%。源自 Alizadeh 等（2000)，经 Nature Publishing Group 允许。(C) **对化疗药物的敏感性**。60个肿瘤细胞系中6817个基因表达谱与232种药物进行相关性分析。该图显示30个肿瘤细胞系（列）的基因表达谱（行）对其中一种药物 cytochalasin D 敏感或耐受。源自 Staunton 等（2001)，经美国国家科学院允许。(D) **乳腺癌临床后果预测**。在98例原发性乳腺癌中获得了25000个基因的表达谱。该图显示78例肿瘤（行）中70个预后标记基因的表达谱（列）。所有患者均接受外科手术和放射治疗；黄线将额外接受化疗（在黄线以上）和不接受化疗的患者分开。数据资料源自 vant't Veer 等（2002)，经 Nature Publishing Group 允许。

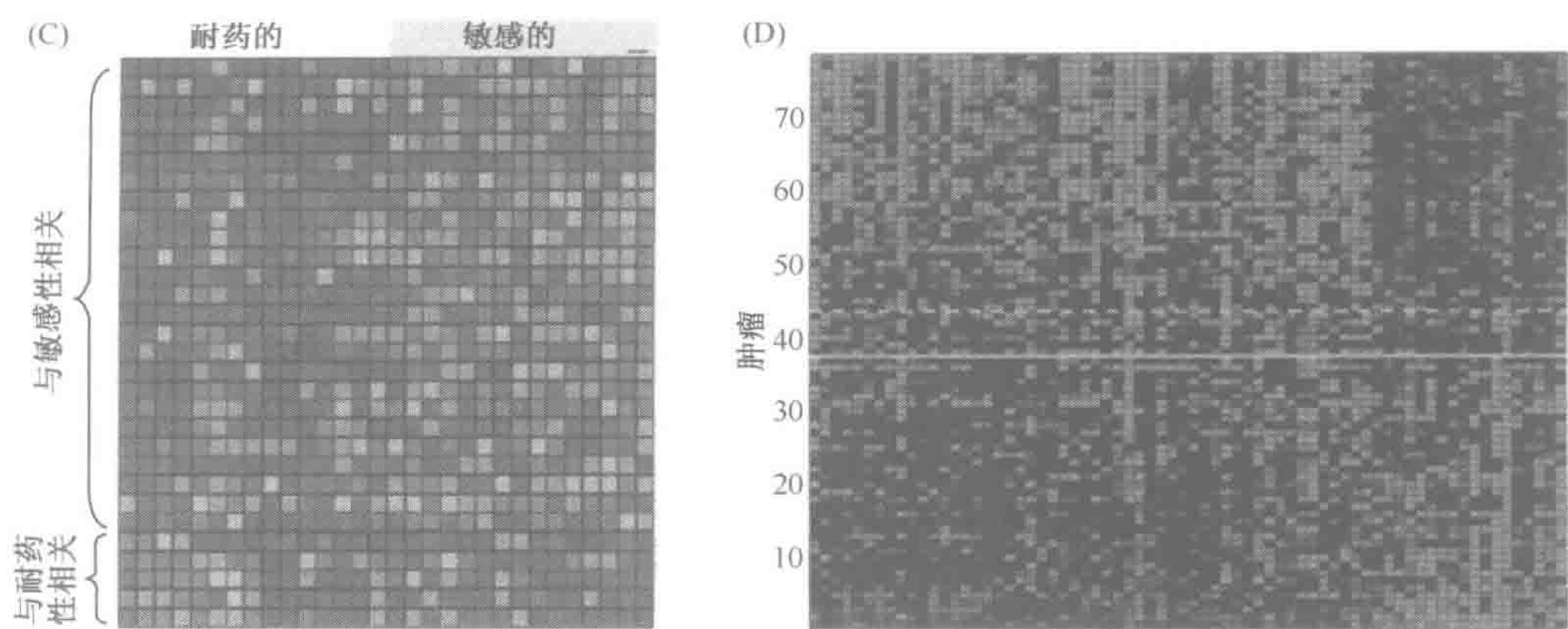


图 17.18（续）

2002)。直到最近，实体肿瘤中变化的复杂性使已经存在的分析方法难以应用，但是，基因**表达谱**（expression profiling）正开始提供有用的亚型分类，而且涉及预后和治疗。图 17.18 举例说明了表达芯片最近在肿瘤中的应用。

▶ 建立了首例基于特异性分子改变信息的治疗方法。ErbB2 受体酪氨酸激酶的一个单克隆抗体 Herceptin，对存在 *ERBB2* 基因扩增的乳腺癌有效，而对无此特征的乳腺癌无效（de Bono and Rowinsky，2002）。Gleevec（Imatinib，ST-1571）是 BCR-ABL 融合蛋白的特异抑制剂，对慢性髓性白血病的治疗效果明显（Savage and Antman，2002）。但愿这些是根据肿瘤的分子变化的信息所设计的第一批抗肿瘤药物。

（富伟能 译）

进一步阅读

Bishop JM (1983) Cellular oncogenes and retroviruses. *Annu. Rev. Biochem.* **52**, 301–354.

Hanahan D, Weinberg R (2000) The hallmarks of cancer. *Cell* **100**, 57–70.

Stanbridge EM (1990) Human tumor suppressor genes. *Annu. Rev. Genet.* **24**, 615–657.

参考文献

Alizadeh AA, Eisen MB, Davis RE *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

Al-Tassan N, Chmiel NH, Maynard J *et al.* (2002) Inherited variants of *MYH* associated with somatic G:C→T:A mutations in colorectal cancer. *Nature Genet.* **30**, 227–232.

Armstrong SA, Staunton JE, Silverman LB *et al.* (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* **30**, 41–47.

Blume-Jensen P, Hunter T (2001) Oncogenic kinase signalling. *Nature* **411**, 355–365.

Cavenee WK, Dryja TP, Phillips RA *et al.* (1983) Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**, 779–784.

Chissoe SL, Bodenteich A, Wang Y-F *et al.* (1995) Sequence and analysis of the human *ABL* gene, the *BCR* gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27**, 67–82.

De Bono JS, Rowinsky EK (2002) The ErbB receptor family: a therapeutic target for cancer. *Trends Mol. Med.* **8**(4 Suppl), S18–S26.

Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767.

Fishel R, Lescoe MK, Rao MRS *et al.* (1993) The human mutator gene homolog *MSH2* and its association with hereditary non-polyposis colon cancer. *Cell* **78**, 539–542.

Fodde R, Smits R, Clevers H (2001) *APC*, signal transduction and genetic instability in colorectal cancer. *Nature Rev. Cancer* **1**, 55–67.

Forozan F, Karhu R, Kononen J *et al.* (1997) Genome screening by comparative genome hybridization. *Trends Genet.* **13**, 405–409.

Futreal PA, Kasprzyk A, Birney E *et al.* (2001) Cancer and genomics. *Nature* **409**, 850–852.

Hahn WC, Weinberg RA (2002) Modelling the molecular circuitry of cancer. *Nature Rev. Cancer* **2**, 331–340.

Harbour JW, Dean DC (2000) The Rb/E2F pathway: expanding roles and emerging paradigms. *Genes Dev.* **14**, 2393–2409.

Hoeijmakers J (2001) Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366–374.

Jallepalli PV, Lengauer C (2001) Chromosome segregation and cancer: cutting through the mystery. *Nature Rev. Cancer* **1**, 109–117.

Jiricny J, Nystrom-Lahti M (2000) Mismatch repair defects in cancer. *Curr. Opin. Genet. Dev.* **10**, 157–161.

Jones PA, Baylin SB (2002) The fundamental role of epigenetic events in cancer. *Nature Rev. Genet.* **3**, 415–428.

Kinzler KW, Vogelstein B (1996) Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170.

Knudson AG (2001) Two genetic hits (more or less) to cancer. *Nature Rev. Cancer* **1**, 157–162.

Lakhani SR, Ashworth A (2001) Microarray and histopathological analysis of tumours: the future and the past? *Nature Rev. Cancer* **1**, 151–157.

Lamlum H, Ilyas M, Rowan A *et al.* (1999) The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. *Nature Med.* **5**, 1071–1075.

Lowy DR, Willumsen BM (1993) Function and regulation of RAS. *Annu. Rev. Biochem.* **62**, 851–891.

Malkin D (1994) Germline p53 mutations and heritable cancer. *Annu. Rev. Genet.* **28**, 4443–4465.

Maser RS, DePinho RA (2002) Connecting chromosomes, crisis and cancer. *Science* **297**, 565–569.

Mitelman F, Johansson B, Mertens F (eds) Mitelman Database of Chromosome Aberrations in Cancer (2002). http://cgap.nci.nih.gov/Chromosomes/Mitelman

Pinkel D (1994) Visualizing tumor amplification. *Nature Genet.* **8**, 107–108.

Pomeranz J, Schreiber-Agus N, Liegois NJ *et al.* (1998) The Ink4a tumor suppressor gene product, p19Arf, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell* **92**, 713–723.

Rabbitts TH (1994) Chromosome translocations in human cancer. *Nature* **372**, 143–149.

Sanchez-Garcia I (1997) Consequences of chromosomal abnormalities in tumor development. *Ann. Rev. Genet.* **31**, 429–453.

Savage DG, Antman KH (2002) Imatinib mesylate – a new oral targeted therapy. *New Engl. J. Med.* **346**, 683–693.

Scully R, Livingston DM (2000) In search of the tumour-suppressor functions of *BRCA1* and *BRCA2*. *Nature* **408**, 429–432.

Sidransky D (2002) Emerging molecular markers of cancer. *Nature Rev. Cancer* **2**, 210–219.

Smith G, Carey FA, Beattie J *et al.* (2002) Mutations in APC, Kirsten-ras, and p53-alternative genetic pathways to colorectal cancer. *Proc. Natl Acad. Sci. USA* **99**, 9433–9438.

Staunton JE, Slonim DK, Coller HA *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98**, 10787–10792.

Su AI, Welsh JB, Sapinoso LM *et al.* (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.* **61**, 7388–7393.

Thiagalingam S, Laken S, Willson JKV *et al.* (2001) Mechanisms underlying losses of heterozygosity in human colorectal cancers. *Proc. Natl Acad. Sci. USA* **98**, 2698–2702.

Tomlinson I, Halford S, Aaltonen L *et al.* (2002) Does MSI-low exist? *J. Pathol.* **197**, 6–13.

Tosi S, Giudici G, Rambaldi A *et al.* (1999) Characterization of the human myeloid leukemia-derived cell line GF-D8 by multiplex fluorescence in situ hybridization, subtelomeric probes and comparative genomic hybridization. *Genes Chrom. Cancer* **24**, 231–241.

van't Veer LJ, Dai H, van de Vijver MJ *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536.

Vousden KH (2000) p53: death star. *Cell* **103**, 691–694.

White RL (1998) Tumor suppressing pathways. *Cell* **92**, 591–592.

Wu CL, Roz L, Sloan P, Read AP *et al.* (1997) Deletion mapping defines three discrete areas of allelic imbalance on chromosome arm 8p in oral and oropharyngeal squamous cell carcinomas. *Genes Chrom. Cancer* **20**, 347–353.

Zhou B-BS, Elledge S (2000) The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439.

第 18 章　个体和群体的遗传检测

本章内容

18.1　概述

遗传学家并非独占以 DNA 为基础的诊断领域。例如，对于微生物学家和病毒学家，PCR 是鉴定病原体的核心工具。血液学家，肿瘤学家和其他病理学家都采用 DNA 检测作为诊断的基础。然而，鉴于本章的目的，我们把遗传检测定义为孟德尔因子的检测。这些因子可能表明一个人罹患或传递某种疾病的风险（可能是或可能不是孟德尔式疾病），或者被用于确定其身份或揭示其与他人的亲属关系。

我们将采用两种最常见的孟德尔式疾病，囊性纤维化（CF）和杜兴肌营养不良（DMD）来说明各种可能的检测方法。这两种疾病都涉及具有广泛的等位基因异质性的大基因。但是除此以外，他们在 DNA 诊断方面会出现相当不同的一系列问题（表 18.1）。这两种疾病共同说明了孟德尔式疾病检测的许多问题。正如本书的一贯风格，我们重在原理而不是实际的细节。对于具体操作感兴趣的读者可以在 http//www.cmgs.org 找到一系列更为常见的孟德尔式疾病“最实用”实验室诊断准则。这些是由英国临床分子遗传协会协议专题讨论会制订的。由 Elles 和 Mountford（进一步阅读）所著书中叙述了许多疾病的检测方法。

表 18.1 囊性纤维化和杜兴肌营养不良的遗传学对比

囊性纤维化	杜兴肌营养不良
常染色体隐性遗传	X 连锁隐性遗传
功能丢失突变	功能丢失突变
相当大的基因	巨型基因
250kb 基因组 DNA	2400bp 基因组 DNA
27 个外显子	79 个外显子
6.5kb mRNA	14kb mRNA
几乎所有的突变为单核苷酸变化	65%的突变为含有一个或多个完整外显子的缺失
	5%突变为重复
	30%突变为无义、剪接位点等的突变
	错义突变非常少见
新突变极其罕见	新突变很常见
嵌合型不是囊性纤维化的问题	嵌合型常见
很少有基因内重组	重组热点（12%位于基因任意一端的标记之间）

囊性纤维化和杜兴肌营养不良基因检测需要几套不同的方法。

当临床医生带一份患者 DNA 来实验室要求做诊断性检测时，实验室可能试图回答三个可能的问题。

▶ 患者是否有能解释其疾病的某种基因突变？无论现在还是将来，这个问题都是不可答复的。基因检测必须具有针对目标。即使将患者整个基因组测序作为诊断手段成为可能，如果没有特定的候选序列名单，我们也不能确定两个人 DNA 序列间几百万个差别中哪一个可能是疾病的原因。

▶ 患者存在可能导致其疾病的某个特定基因突变吗？回答这个问题的标准方法在节 18.3 中述及。而另一种方法，基因示踪将在节 18.5 中述及。

▶ 患者有 *CFTR* 基因第 508 位编码苯丙氨酸密码子的三碱基缺失吗？这类问题在什么情况下可能被提出以及回答该问题的方式将在节 18.4 中述及。

18.2 受检材料的选择：DNA，RNA 或蛋白质

遗传检测几乎总是由 PCR 完成，采用的方法在节 5.2 中叙述。Southern 印迹杂交的几个应用包括检测主要基因重排或断裂以及检测脆性 X 染色体病和肌强直营养不良的全突变（节 16.6.4）。PCR 的灵敏性允许我们采用广泛的组织样品，包括：

▶ **血液样品**：最广泛采用的成人 DNA 的来源；

▶ **漱口液或颊膜刮片**：由于非损伤性，他们特别适合群体筛查计划。漱口液可获得足够几十次检测用的 DNA。通过全基因组扩增（节 5.2.4），对单一样品的更广泛的检测也许成为可能；

▶ **绒毛膜绒毛活组织检查样品**：胎儿 DNA 的最佳来源（优于羊膜腔穿刺术样品）；

▶ **取自 8 细胞期胚胎的 1 个或 2 个细胞**：用于体外受精后植入前诊断；

- **毛发、精液等**：用于犯罪调查；
- **归档的病理学标本**：在没有保存 DNA 时，用于死者分型或检测肿瘤遗传学改变。只有短序列，250bp 或更小，才可以从固定的组织样品中可靠地扩增；
- **Guthrie 卡片**：卡片上有干燥血斑送实验室待检，在英国和其他地方用于筛查新生儿苯丙酮尿症（PKU）。并不是全部血斑都用于筛查检测，所以如果卡片得以保存，这可能作为死亡儿童 DNA 的来源。

RNA 优于 DNA，但更难以获得与操作

如果一个基因必须被筛查未知的突变，用 RT-PCR（节 5.2.1）进行检测有几项优点。DNA 检测通常涉及分别扩增和检测每个外显子，对于含有许多外显子的基因，这可能是一个较繁琐的工作。大多数突变筛查方法（表 18.2）所能筛查的片段长度大于外显子平均长度。因此，通过更少的反应就可以完成 RT-PCR 产物的检测。此外，只有 RT-PCR 能可靠地检测出异常剪接，这种剪接通常很难从 DNA 序列变化进行预测，它可能是由内含子深处隐蔽剪接位点的激活所引起（节 16.4.1）。然而，RNA 非常不便于获得与操作。为了避免 mRNA 降解，样品必须被极其小心地处理及快速地操作。目的基因可能在易于获得的组织中不表达。此外，许多突变导致不稳定的 mRNA（节 11.4.4，节 16.4.1），因此，杂合个体的 RT-PCR 产物可能只显示正常等位基因。

表 18.2　筛查基因突变的方法

该表总结了常规诊断检测采用的每种方法的优点和缺点。详见节 18.3

方法	优点	缺点
Southern 印迹，cDNA 探针杂交	检测大的缺失和重排的唯一方法	费力，昂贵，需几 μg DNA
测序	检测全部改变 突变可完全定性	昂贵
杂合双链凝胶泳动度	非常简单 价廉	仅适用于小于 200bp 序列
		灵敏度有限，不能显示突变位置
dHPLC	迅速　高通量　定量	设备昂贵不能显示突变位置
SSCP	简单　价廉	仅适用于小于 200bp 序列，不能显示突变位置
DGGE	高灵　敏度	引物的选择很关键，引物昂贵，不能显示突变位置
错配化学切割法	高灵敏度 能显示突变位置	有毒化学物质 实验困难
PTT	对链终止突变，其灵敏度高	仅对链终止突变，昂贵，技术困难
	能显示突变位置	通常需要 RNA
定量 PCR	检测杂合子缺失	昂贵
芯片	迅速 高通量	昂贵 待检基因范围有限
	可能检测到和确定全部改变	

dHPLC，变性高效液相色谱；SSCP，单链构象多态性；DGGE 变性梯度凝胶电泳；PTT 蛋白截短实验。

蛋白质功能测定在基因检测中发挥作用

基于蛋白质的功能测定可将高度异质的等位基因产物分成简单的两组：功能性和非功能性，毕竟这是大多数诊断中基本的问题。功能检测的问题是他们对某一特殊蛋白质是特异的，相反，DNA 技术是一般性的，这对于诊断性实验室有明显的优点，此外，这也能促进技术发展，因为任何新技术可用于解决许多问题。

18.3 筛查基因突变

绝大部分疾病，如囊性纤维化和杜兴肌营养不良，都存在广泛的等位基因异质性（节 16.3.2）。因此，诊断性检测通常涉及在一个或多个相关基因内或附近的任何区域寻找突变。表 18.2 列出了许多可以采用的方法，这些方法将在下文简要叙述。对于实验细节，参见 Cotton，Edkins 和 Forrest，或 Elles 和 Mountford 所著的书（进一步阅读）。

筛查一个大组患者某候选基因全部序列变异将发现许多不同的变异体，确定一个变异体是否为致病性是非常困难的（如果为致病性，即是待寻找的突变）。框 16.5 给出了一些判定的原则。

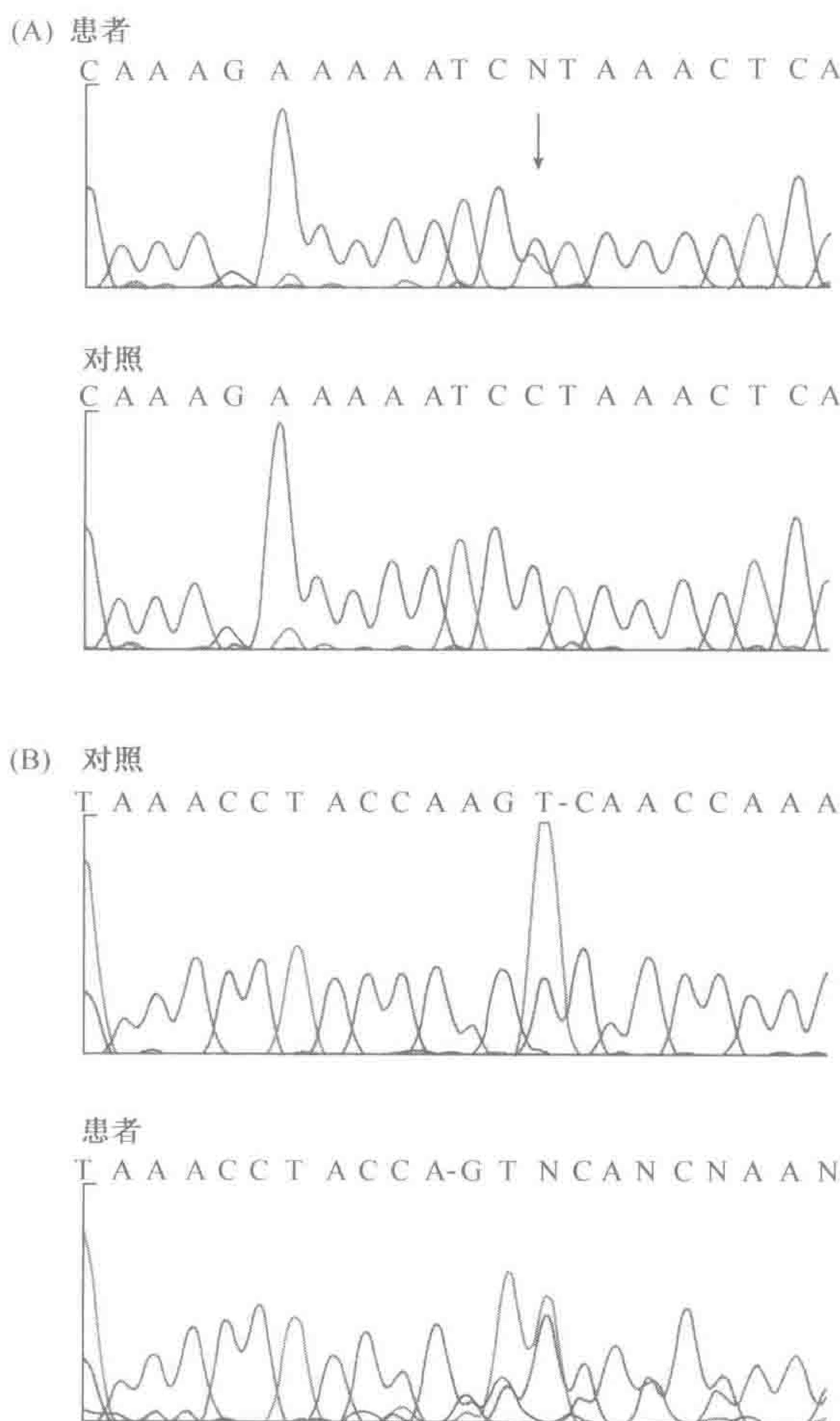


图 18.1 采用测序的突变检测

(A) 在外显子 3 的一个碱基替换。双峰（箭头）表示杂合子突变 g. 332C→T（p. P67L）(B) 外显子 19 的单碱基缺失 3659delC。缺失下游序列混杂，反映出该杂合子中两个等位基因的重叠序列。突变可通过对反义链测序予以证实。英国 Manchester St Mary 医院 Andrew Wallace 博士惠赠。

18.3.1 基于测序的方法

目前自动荧光测序仪是标准的实验室设备。测序作为筛查突变的主要方法，越来越有吸引力（图 18.1），其他的筛查方法主要通过确定哪一个扩增子需要测序来减轻测序的工作量。由于测序已变得更加便宜与简便，减轻工作量的需要因此已减小了。以前使用的其他方法已不受欢迎。现在主要采用不需要直接测序的其他方法，因为这些方法快捷（变性高效液相色

谱；dHPLC)、便宜（单链构象多态性；SSCP)，或者能提供一些特殊的信息（蛋白截短实验；PTT 和定量 PCR)。

测序比其他方法能提供更多的数据，因此对数据分析的需求增加了。只要序列质量好，现有程序就可以自动报告待测序列与标准序列之间的差别。序列质量是十分关键的，它能避免人为假象并可靠地检测出杂合子中的碱基替换。

为了对基因组 DNA 进行测序，通常分别扩增每个外显子，每个外显子需与大约 20bp 的旁侧内含子一起扩增。这种做法效率低，因为人类基因平均外显子的长度 (145bp) 远远低于一次优质测序所能读出的序列长度（500～800bp)。可能的解决方法是使用 RT-PCR 或使用外显子连锁 PCR（meta-PCR，Wallace *et al.*，1999；图 18.2)。基于微阵列的再测序在下面介绍（节 18.4.1)。

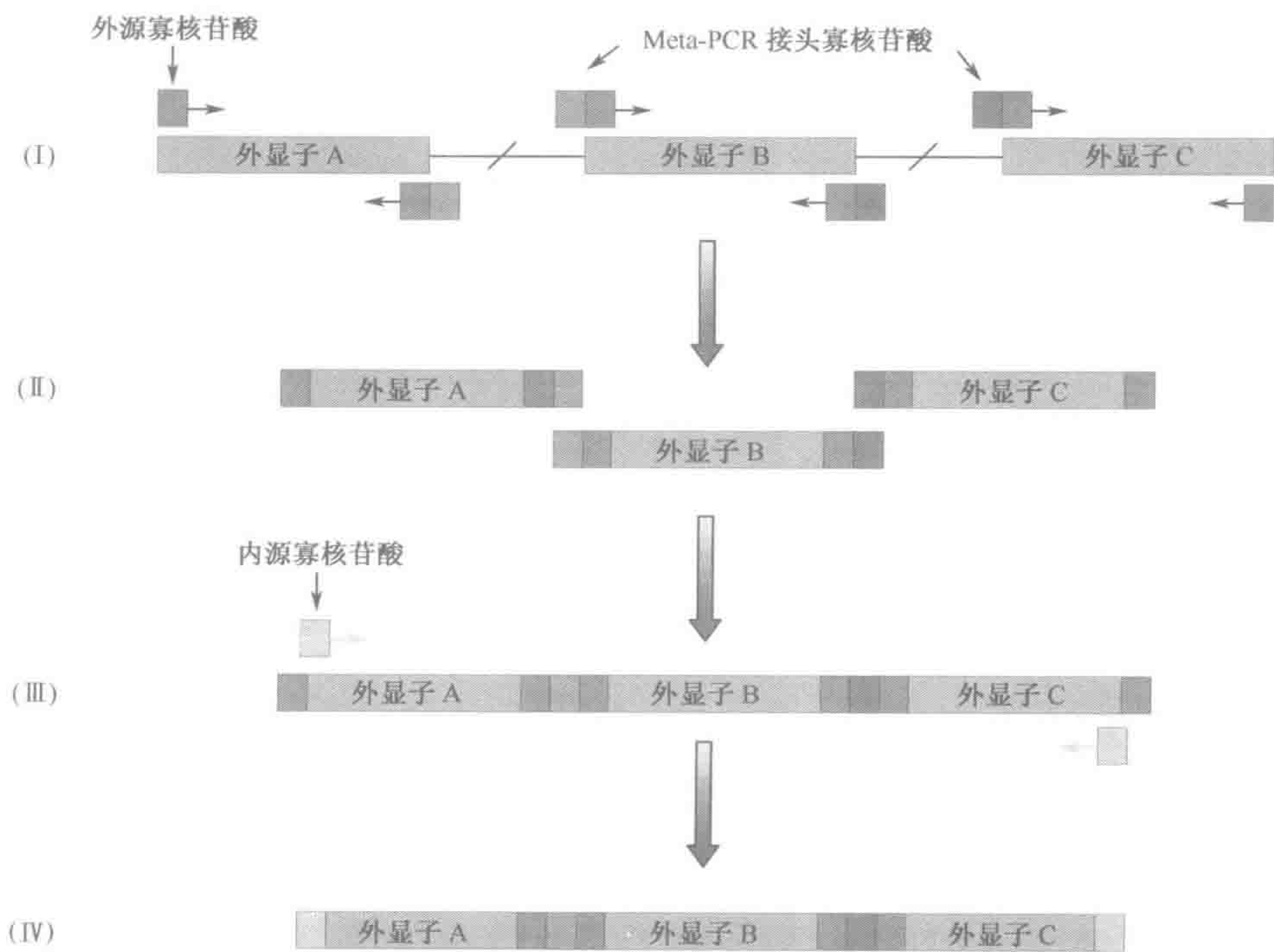


图 18.2　外显子连锁 PCR 的原理（Meta-PCR)

这是一项用于克服人类外显子平均长度（145bp）与测序产物最佳长度（500～800bp）差异的技术，同时保留了研究基因组 DNA 的优点。(Ⅰ) 2～5 个外显子（含有 20bp 旁侧内含子，为清晰起见未画出）采用在 5′端携带特异匹配接头子的引物进行 PCR 扩增。(Ⅱ) 引物消耗尽后，多联体由接头子指导形成。(Ⅲ) 采用特异设计的引物对预计的多联体进行第二轮 PCR 扩增。(Ⅳ) 产物可以设计成外显子以任意顺序组合的形式。见 Wallace Wu 和 Elles (1999)。

18.3.2　基于检测错配或杂合双链的方法

许多检测利用杂合双链的特性检测两条序列间的差异。大多数突变是以杂合形式出现（即使对常染色体隐性疾病，非血缘双亲所生受累个体很可能是复合杂合子，携带两

种不同的突变)。将杂合性的受试 PCR 产物加热使之变性，然后缓慢降温即可形成杂合双链。为了检测纯合性突变或男性 X 连锁突变，必须要加入一些野生型 DNA 作为参照，可以利用杂合双链的几种特性：

- 在非变性聚丙烯酰胺凝胶中，杂合双链通常有异常的泳动度（图 18.3 A，下部），应用特殊的凝胶（Hydrolink™，MDE™）可望提高分辨率，这是个可使用的特别简便的方法。若待测片段长度不超过 200bp，则可以检测出插入、缺失和大部分但不是全部的单碱基替换（Keen *et al.*，1991）。

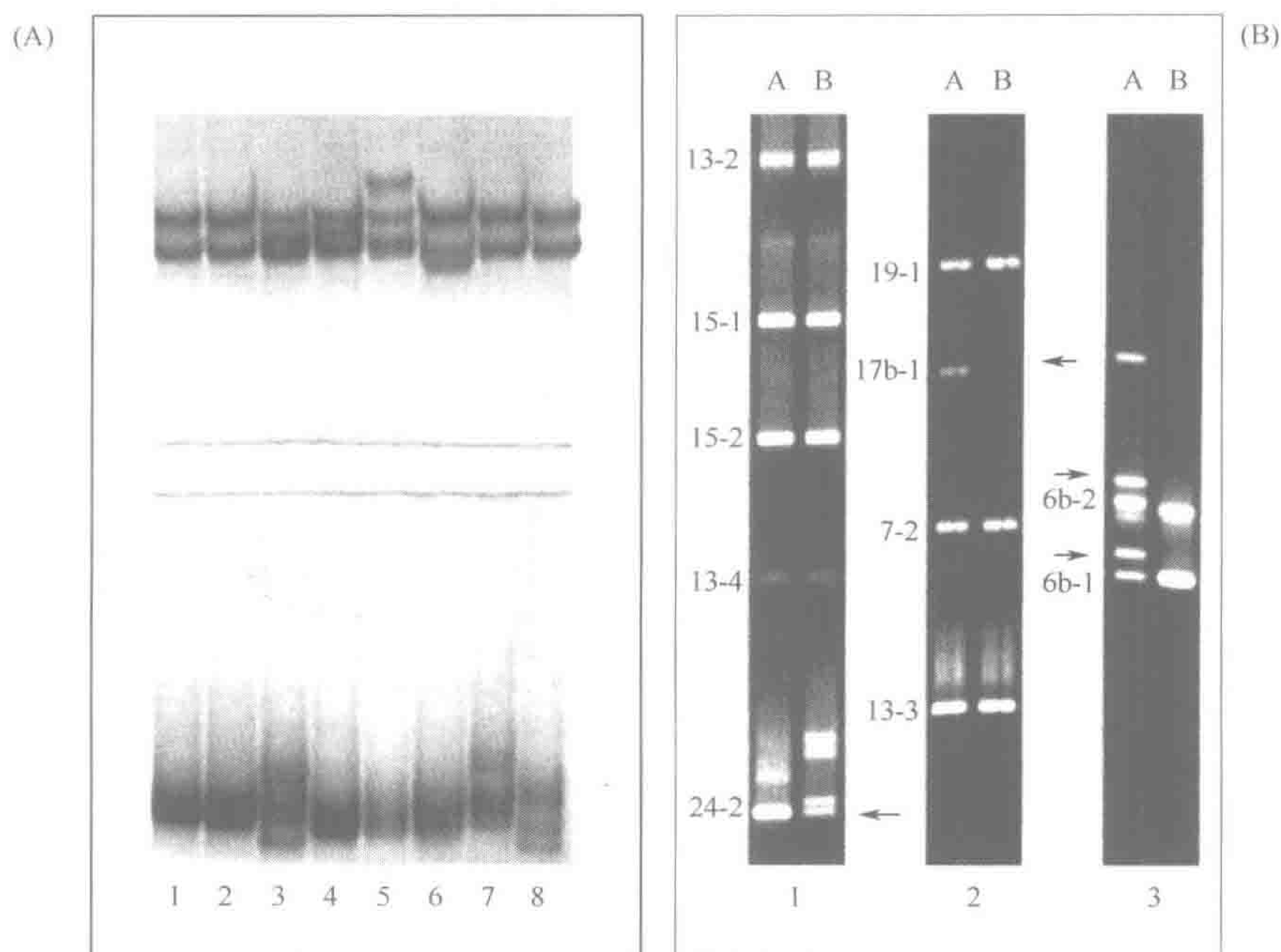


图 18.3 筛查 *CFTR* 基因突变

(A) **杂合双链和 SSCP 分析**。将外显子 3 从基因组 DNA 经 PCR 扩增。变性后，样品上样到非变性聚丙烯酰胺凝胶，将每种产物中的一部分重新退火以产生双链 DNA，在凝胶中其泳动速度快一些，使杂合双链条带见于泳道底部，在同样凝胶中，单链 DNA 泳动很慢（在泳道上部），泳道 1 和 2 显示的是典型的野生型序列模式，变异的模式见于泳道 3～8，测序表明分别为 G85E，L88S，R75X，P67L，E60X，R75Q 的突变。Manchester St Mary 医院 Andrew Wallace 博士惠赠。(B) **变性梯度凝胶电泳**。*CFTR* 基因外显子一个片段或多个片段经 PCR 扩增，在含有梯度尿素-甲醛变性剂的 9%聚丙烯酰胺凝胶中进行电泳。含有杂合变异体的扩增子条带通常分离成四条亚带（箭头）。在显示的泳道中，患者 A（在每一组的左侧泳道）在扩增子 6 有变异体，患者 B（右侧泳道）在扩增子 17 和 24 有变异体。变异体的性质表明患者 B 是 R1070Q（外显子 17）杂合子。其他的变异体是非致病性的。荷兰 Groningen 大学 Hans Scheffer 博士惠赠。

- 杂合双链有异常的变性谱。**变性高效液相色谱**（denaturing high performance liquid chromatography，dHPLC，图 18.4A）和**变性梯度凝胶电泳**（denaturing gradient gel electrophoresis，DGGE，图 18.3B）即利用了这一特性。在这两种检测方法中，片段变性时，其泳动变化明显。这些方法需要调整以便于检测特殊的 DNA 序列，因此

最适合对多样品中的指定片段做常规分析。dHPLC 具有高通量的特点，很适合此项应用。DGGE 需要具有 5′多聚（G；C）延伸（GC 夹）的特殊引物-见 Sheffield 等.（1992）。一旦条件优化后，这些方法具有很高的灵敏度。

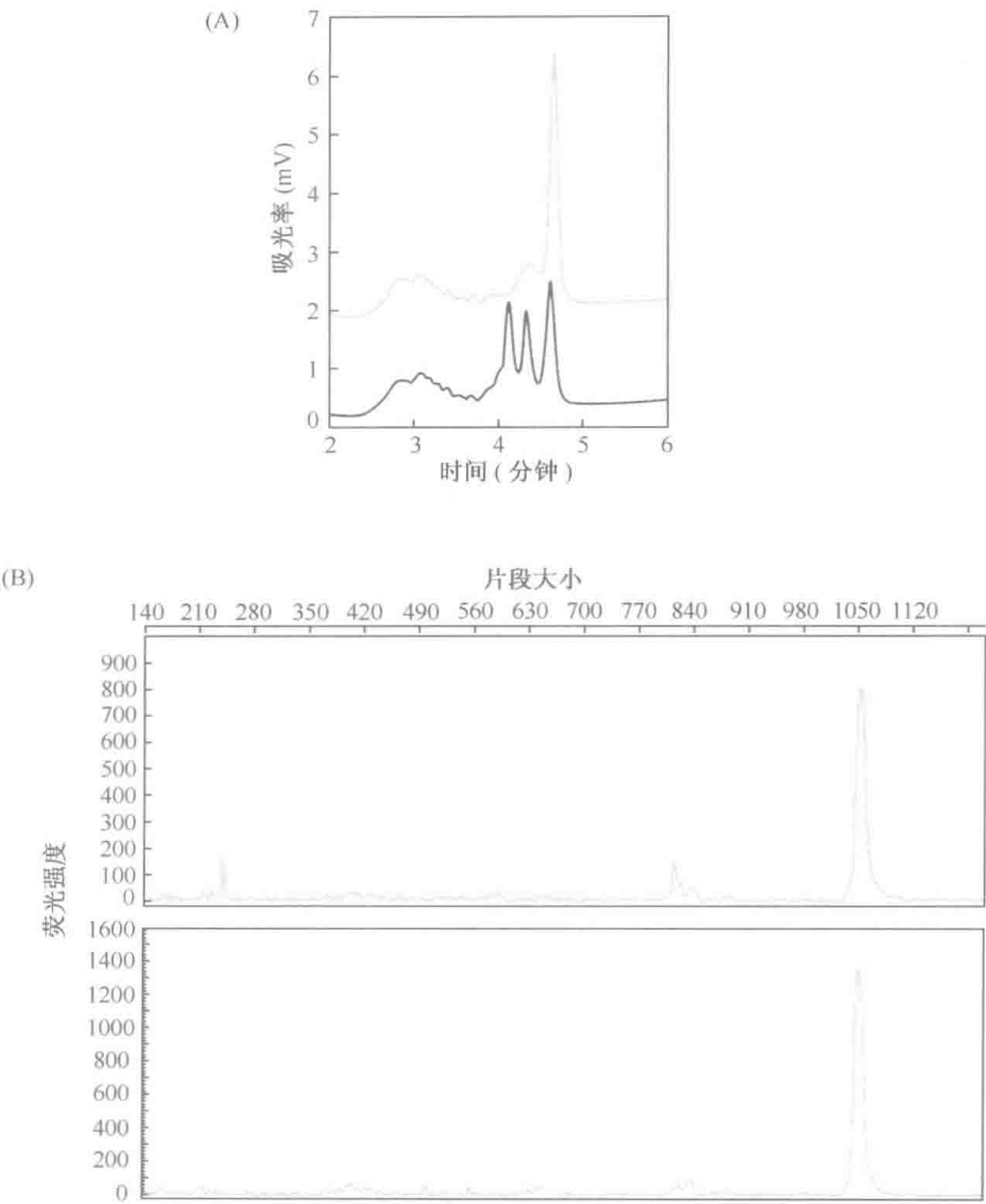


图 18.4　突变筛查

（A）变性高效液相色谱（dHPLC）筛查 DMD 突变。肌细胞增强蛋白基因的外显子 6 在受累男性（黑色轨迹线）和正常对照（浅灰轨迹线）有不同的模式。测序显示剪接位点突变 738＋1G→T。因为这是一个 X 连锁的疾病，对于男性，待检 DNA 必须与同等量正常 DNA 混合，以促成杂合双链的形成。美国马塞诸赛州波士顿儿童医院 Richard Bennett 博士惠赠。（B）错配化学切割筛查突变。含有 *NF2* 基因外显子 6～10 的荧光标记 meta-PCR 产物，上部轨迹图：患者样品；羟胺裂解 1032bp 的 meta-PCR 产物变成 813＋329bp 的片段，这可以显示杂合性内含子 6 剪接突变 600-3c→g。下部轨迹图：对照样品。英国 Manchester St Mary 医院 Andrew Wallace 博士惠赠。

▶ 杂合双链中的错配碱基对化学物质或酶的裂解作用敏感。**错配化学切割法**（chemical cleavage of mismatch，CCM）（图 18.4B）是检测突变的一种敏感的方法。其优点在

于可分析较大片段（长度大于 1Kb），并且能通过产生的片段的大小精确定位错配的位置。然而，这种方法使用剧毒的化学物质，尤其是四氧化锇（尽管可以用高锰酸钾替代），且实验操作相当困难。一种替代方法就是**酶错配切割法**（enzymatic cleavage of mismatch），它使用一些酶，如 T4 噬菌体解离酶或内切酶Ⅶ，不使用有毒化学物质，能得到相同的结果。使人遗憾的是，对于大多数人操作而言，制胶的质量还应得到进一步加强。

在所有这些方法中，现在仅有 dHPLC 在主要的诊断性实验室中被广泛应用。

18.3.3 基于单链构象分析的方法

单链 DNA 有折叠并形成复杂结构的倾向，复杂结构通过分子内弱键，特别是碱基配对的氢键维持其稳定。在非变性凝胶中，这种结构的电泳泳动不仅取决于链的长度，还取决于其构象，这种构象则由 DNA 序列决定。SSCP 是通过扩增 DNA 样品（可能是 RT-PCR 产物），变性，短促降温，上样于非变性聚丙烯酰胺凝胶（图 18.3A）而被检测的。引物可以使用放射性物质标记，或未标记产物可以通过银染检测，所观察到的精确带型很大程度上取决于条件的细节。对照样品必须经电泳分离，因此，可以观察到与野生型的差别。SSCP 非常便宜，对于大到 200bp 长的片段（Sheffield *et al.*，1993），灵敏度尚好（大约 80%）。所以，SSCP 仍得到广泛的应用。如图 18.3A 所示，SSCP 和杂合双链分析可以结合在同一块凝胶上进行。SSCP 的进一步完善，即**双脱氧指纹谱**（dideoxy fingerprinting），借助 SSCP 可分析序列梯状电泳条带中的每一个条带，据称灵敏度可达到 100%（Sarkar *et al.*，1992）。

18.3.4 基于翻译的方法：蛋白截短实验

PTT（图 18.5）是一种特异性的检测方法，用于检测移码突变、剪接位点突变或产生提前终止密码子的无义突变（van der Luijt *et al.*，1994）。研究起始材料是 RT-PCR 产物，或偶尔采用基因组 DNA 中大的单个外显子，如 *APC* 基因 6.5kb 长的外显子 15 或 *BRCA1* 基因 3.4kb 长的外显子 10。正常情况下防止形成截短蛋白的无义介导 mRNA 降解（节 16.4.1）不会发生，因为在检测中没有外显子-外显子剪接。很明显，PTT 的强项与不足是它只能检测一定类型的突变，这对囊性纤维化的诊断不会有什么帮助，因为囊性纤维化中大多数突变是非截短性的。但是在杜兴肌营养不良、腺瘤息肉或是 *BRCA1* 相关乳腺癌中，错义突变不常见，发现的这类变化很可能是巧合的并且是非致病性的。对于这类疾病，PTT 有几项优点，它忽略了沉默或错义碱基替换，并且（类似错配切割法，但与 SSCP 不同）它显示了突变的大致位置。人们发展了几种略微改进的方法以得到更为清晰的结果，通常加入免疫沉淀步骤，但 PTT 中仍是一个难以完善操作的技术。

18.3.5 检测缺失的方法

纯合性或半合性缺失易于检测：缺失的序列将不被 PCR 扩增。考虑到不能扩增的其他解释是重要的：可能存在 PCR 技术的失误或引物的结合位点中存在一个碱基替换。可以通过更换引物或使用 Southern 印迹来验证缺失的存在。DMD 突变中约 60%为一

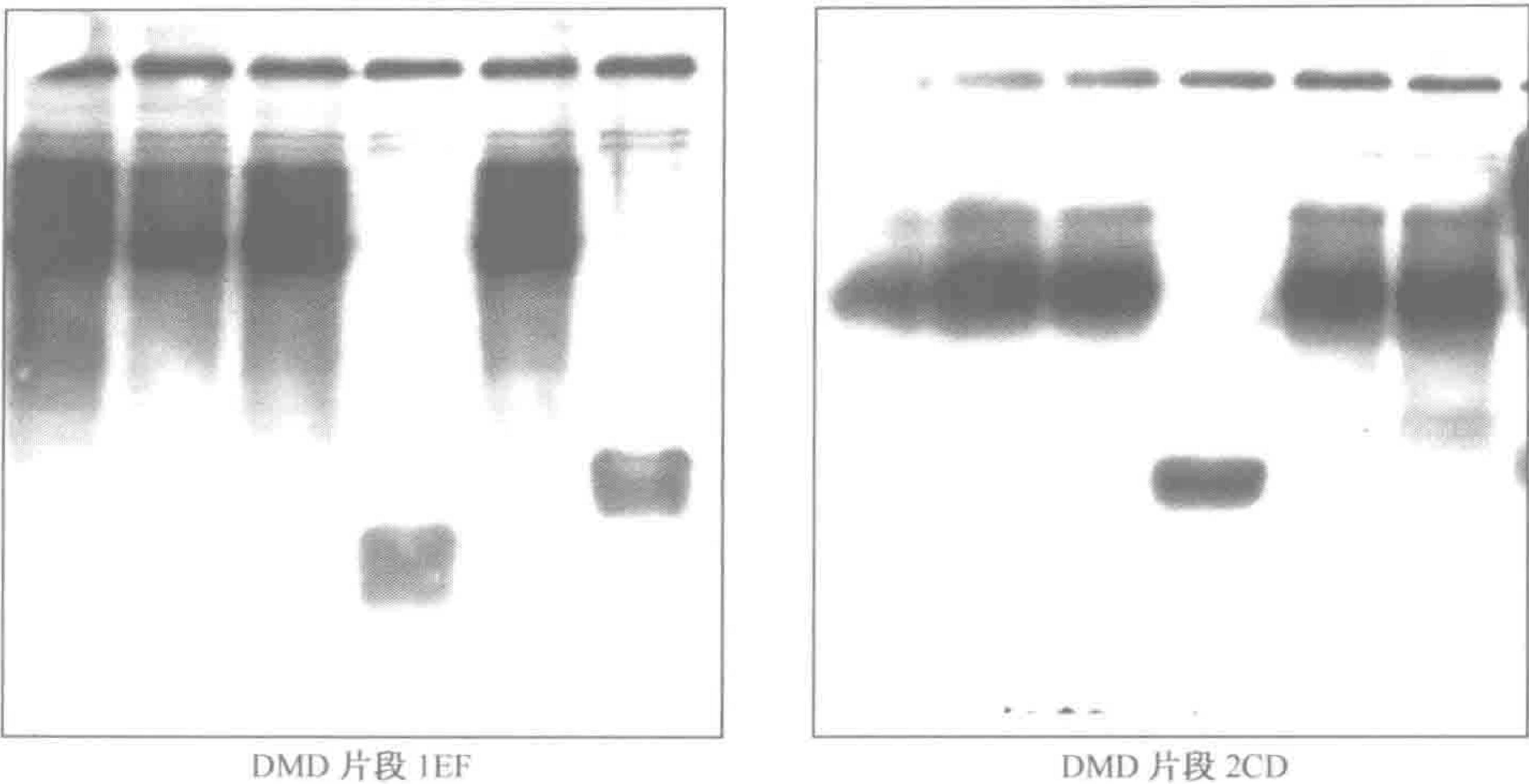


图 18.5　蛋白截短实验（PTT）用于 DMD 突变筛查

偶联的转录-翻译反应用于制备标记的由 mRNA 片段编码的多肽产物。含有提前中止密码子的片段会产生截短的多肽。在这里，RT-PCR 用于扫描一系列 DMD 患者整个抗肌萎缩蛋白基因的十个互相重叠的片段。泳动速度快的多肽代表含有终止密码的片段，异常产物的大小表明了该片段中终止密码子的位置。图片由 J. T. den Dunnen 和 D. Verbove（Leiden，荷兰）博士惠赠。详见http://www.dmd.nl。

个或多个外显子的缺失（图 16.3）。对于受累男性，两个多重 PCR 反应（如图 18.6 所示和基因 5′端的一个待检外显子）可检测出所有缺失中的 98%。大多数缺失丢掉一个以上的外显子。影响非邻近外显子和仅仅单个外显子的缺失需进一步加以验证。

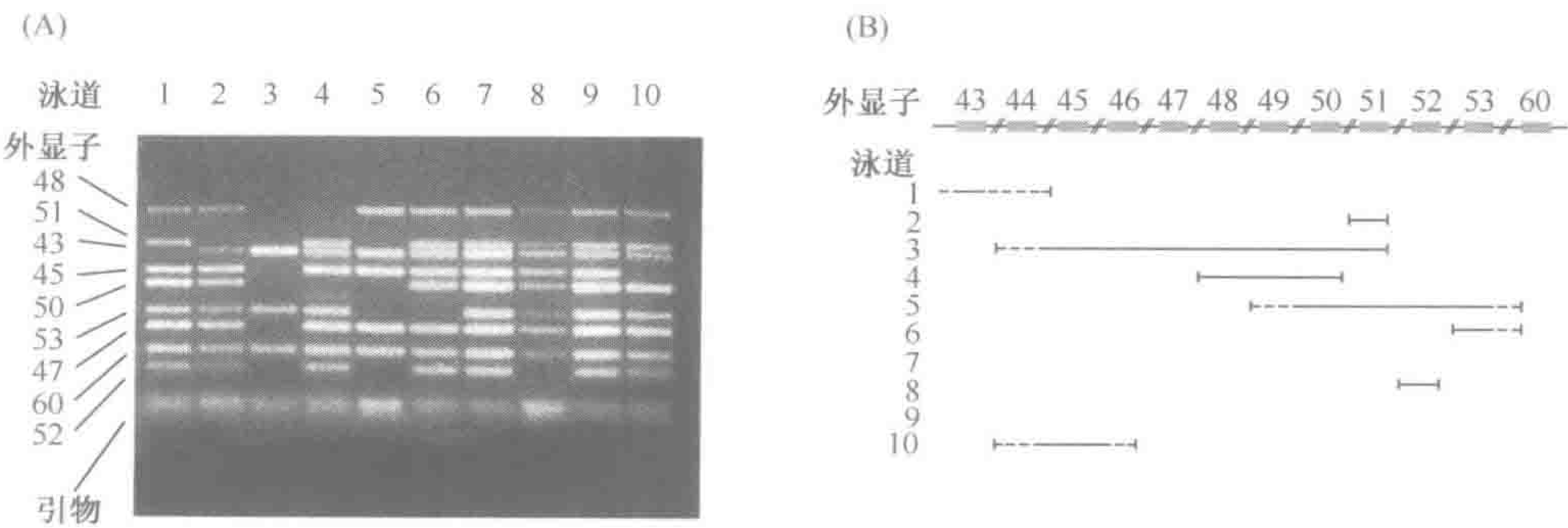


图 18.6　多重筛查男性抗肌萎缩蛋白基因缺失

（A）取样于十个无亲缘关系的杜氏/贝克肌营养不良男孩，对九个外显子进行多重 PCR 扩增的产物。设计 PCR 引物使得每个含有某些旁侧内含子序列的外显子会扩增出不同大小的 PCR 产物。英国利物浦女子医院 R. Mountford 博士惠赠。（B）翻译：实线代表外显子确定的缺失，点线代表涉及到未检测外显子的缺失的可能程度，样品 7 和 9 无缺失，这些患者可能有未在本试验中检测出的点突变或外显子缺失。外显子大小和间距未按比例显示，可与图 16.3 比较。

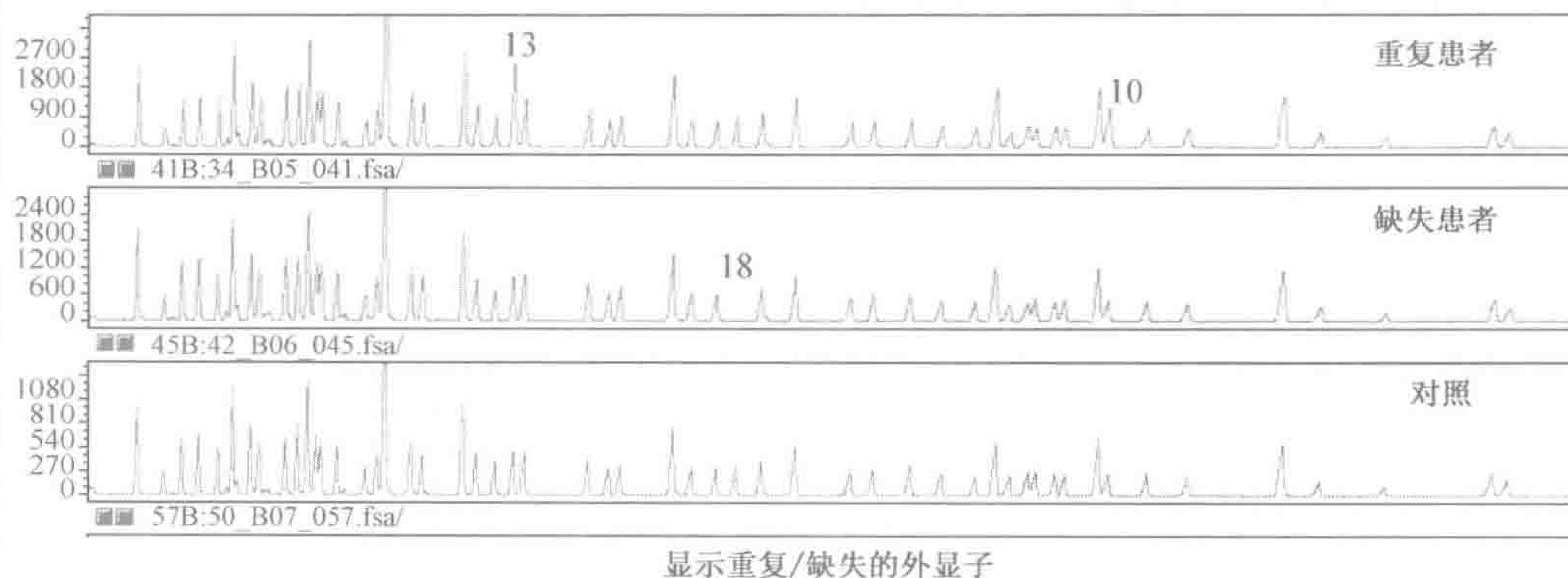
检测女性是否携带抗肌萎缩蛋白基因缺失总体上讲更为困难，这说明了一个或多个完整外显子杂合性缺失引出的特殊问题。当基因组 DNA 逐个扩增外显子，然后用测

序、异源双链分析或 SSCP 进行检测时，这类缺失是检测不到的，因为突变的等位基因没有产物。通过 FISH 或芯片-CGH（图 17.3），可诊断 Mb 大小的片段缺失。对于外显子大小的缺失，需要采用其他方法。测序或对 RT-PCR 产物进行 PTT 分析，除了无法检测整个基因缺失外，可检测其他类型的缺失，但是无义介导的 mRNA 降解可使突变转录物不可见，而且在诊断的条件下，以 RNA 为基础的方法并不总是容易实施。

定量 PCR 的几种体系可用于检测基因组 DNA 杂合性缺失。反应限定在早期指数增长阶段（图 5.2），此时，产物的量反映模板的量。借助荧光来实时测定产物的累积量，并与内在的或平行标准进行比较。荧光信号可由双链 DNA 特异性结合染料产生，如 SYBR® 绿，或由染料和淬灭剂（TaqMan ® assay）标记的序列特异的寡核苷酸切割产生。替代实时 PCR 的一种方法是 Armour（2000)(框 18.1）提出的 MAPH 方法（多重扩增探针杂交）。MAPH 具有高度多重性（至少 50 个短序列的工作量可以在一次实验中被比较），不需要特殊标记的探针或昂贵的仪器，十分适合在中等实验室开展。

框 18.1 多重扩增探针杂交（MAPH）

这是一个比较若干序列基因剂量非常通用的方法，患者基因组 DNA 点样到一个微小（2×3mm）尼龙滤膜上，与 40 个探针的混合物杂交，每个探针与抗肌萎缩蛋白基因的一个外显子特异结合。经仔细冲洗后，滤膜放入 PCR 反应混合物中，该混合物含有与每个探针末端序列结合的一对引物，一个引物可以用荧光标记，可以由基因测序仪分析。设计探针以使外显子可通过 PCR 产物的长度加以区分，并通过荧光测序仪的峰大小加以定量，或通过人工凝胶上条带的相对强度加以定量。因为所有的探针扩增采用同一对引物，困扰多重 PCR 的非均等扩增问题得到避免。MAPH 详细内容，见 Armour 等.（2000）和 MAPH 网址http：www. nott. ac. uk/～pdzjala/maph/maph. html



采用 MAPH 检测抗肌萎缩蛋白基因外显子的缺失和重复

与对照扫描图比较峰大小，很容易看到患者在上部扫描图外显子 10，13 重复（可假定为外显子 11 和 12，在本多重体系中未检测），在中央扫描图有外显子 18 缺失。荷兰 J. T. den Dunnen 和 S. White Leiden 博士惠赠图像。详细内容见http://www. dmd. nl

如果缺失在一个家系中分离，以缺失内的微卫星作图对女性进行分型可能显示明显的非母源性（nonmaternity），即由于缺失，母亲没有将标记等位基因传递给女儿（图 18.7）。在这样的家系中，“非母源性”证明某个女性为携带者，而杂合性（缺失携带者

的女儿或姐妹）证明一个女性不是携带者。在缺失热点的内含子中已识别了几种适合这种研究的标记。在受累男性的缺失首先被确定的家系中，这种方法效果最好。

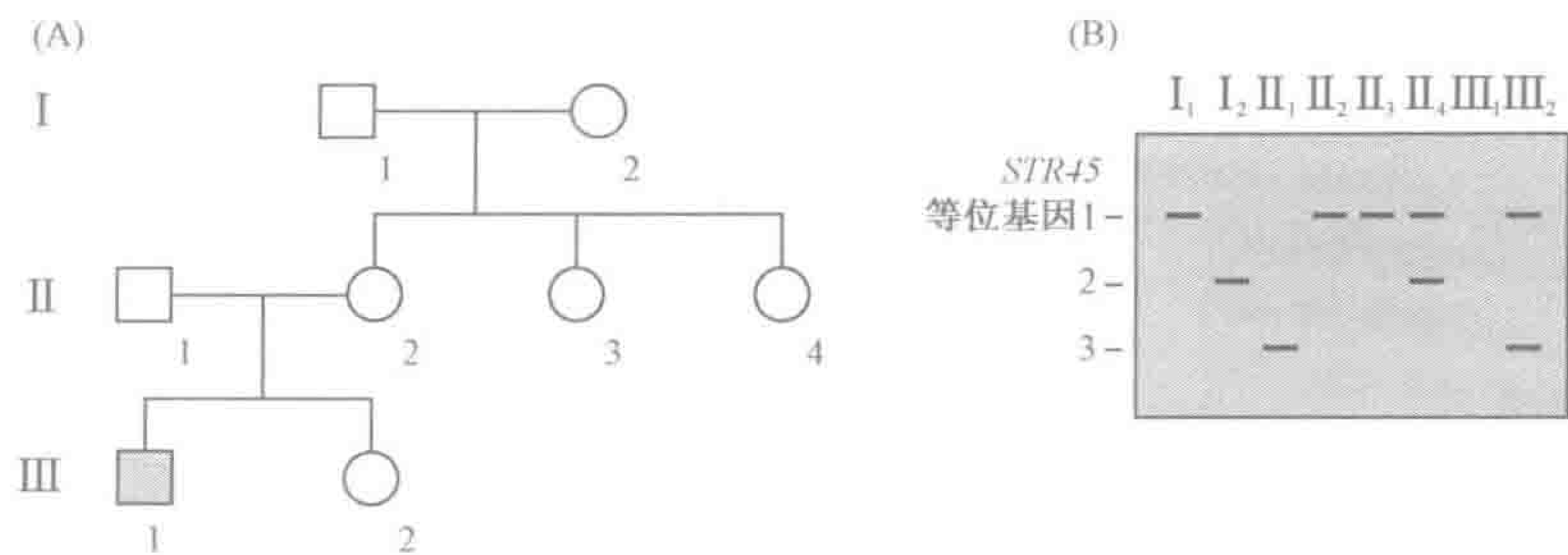


图 18.7　明显的非母源性显示 DMD 家系的缺失携带者

(A) 系谱；(B) 采用基因内标记 *STR45* 基因分型的结果受累男孩Ⅲ-1 有包括 STR45 标记的缺失（凝胶的第 7 道空白），他的母亲Ⅱ-2 和他的姨Ⅱ-3 没有从她们母亲Ⅰ-2 那里遗传 *STR45* 的等位基因，表明缺失在家系中传递。Ⅰ-2 表面上是该高度多态标记的纯合子（泳道 2），但实际上，她是半合子，另一个姨Ⅱ4 和他的妹妹Ⅲ-2 是该标记的杂合子，因此没有携带该缺失。

18.3.6　检测 DNA 甲基化模式的方法

在节 10.4.2 中，我们已经描述了 DNA 甲基化在控制基因表达中的重要性。CpG 二核苷酸的过度或缺乏甲基化是癌症（节 17.4.3）和印记基因表达（节 16.4.4）的一种普遍致病机制。PCR 产物总是非甲基化的，所以至今没有一种方法能够在 DNA 样品中给出甲基化模式的信息。两种主要的方法用于检查甲基化。

▶ 限制酶消化：*Hpa* Ⅱ 仅能切割非甲基化的 CCGG，而 *Msp* Ⅰ 则可以切割任何 CCGG，不论其甲基化与否。因此，甲基化的 CpG 位点使这两种酶产生不同的限制片段。相对而言，含有 CCGG 的 PCR 模板在扩增前可被 *Hpa* Ⅱ 消化，如果该位点未发生甲基化，则模板被切割，没有 PCR 产物产生。

▶ 亚硫酸盐测序（Thomassin *et al.*，1999）当单链 DNA 被亚硫酸钠处理时，胞嘧啶而不是 5-甲基胞嘧啶（5-MeC）被转化为尿嘧啶。亚硫酸盐处理后，DNA 被 PCR 扩增（采用与修饰序列匹配的引物），然后用于常规测序。原样品中胞嘧啶和 5-甲基胞嘧啶分别在 PCR 产物序列中表现为胸腺嘧啶和胞嘧啶。这个方法需要非常精确的对照以给出可靠的数据。

18.4　检测特定序列的变化

对一种已知序列改变存在与否的检测不同于筛查一个基因的突变，相对也更简单。样品总是可以通过测序确定其基因型，但如果仅检测一个核苷酸位点，那么传统的测序不是效率高的方法。表 18.3 总结了一些主要的基因分型方法。许多略微变化的这些和其他方法已由生物技术公司开发成试剂盒。典型的应用包括：

▶ 有限的等位基因异质性疾病的诊断（表 18.4）；

表 18.3 检测特定突变的方法

方法	注释
PCR 扩增的 DNA 经限制酶消化；检测凝胶中产物大小	只有当突变产生自然限制性酶切位点，或使其消失(图 7.6)或采用特殊 PCR 引物制造一个自然限制酶切位点时(图 18.8)，该方法才有效
将 PCR 扩增的 DNA 与等位基因特异寡核苷酸(ASO) 在斑点印记或基因芯片上杂交	特异点突变的一般方法 大规模阵列可筛查几乎任何突变
采用等位基因特异引物 PCR(ARMS 实验)	点突变的一般方法，引物设计关键(图 18.10) 可与芯片技术结合 采用 TaqMan 技术，可提供实时定量
寡核苷酸连接分析(OLA)	检测特定点突变的一般方法(图 18.11)
PCR，引物位于易位断点任意一侧	扩增成功表明存在可疑的缺失或特定的重排
检测扩张重复大小	动态重复性疾病(节 16.6.4) 大的扩张需要 Southern 印迹，小的扩张仅需要 PCR 完成
焦磷酸测序	高通量方法(框 18.2)
SNAPShot	通过引物延伸的小测序(节 18.4.1)
质谱	高通量方法(框 18.2)

表 18.4 有限突变疾病的例子

对于为什么一些疾病显示出有限范围的突变，而其他疾病具有广泛的等位基因异质性的原因，请见节 16.3 的进一步讨论。

疾病	原因	注释
镰状细胞病	只有这种特定的突变导致镰状细胞表型	*HBB* 基因 p.E6V，见图 6.11
软骨发育不全	只有 G380R 导致这种特殊表型；非常高的突变率	两种不同的变化都引起 *FGFR3* 基因 p.G380R(图 16.9)
亨廷顿病 肌强直性营养不良	功能获得突变	不稳定扩张重复 见 16.6.4 节
脆性 X 染色体病	普遍分子机制：不稳定重复的扩张	见节 16.6.4；有其他突变，但罕见
Charcot-Marie-Tooth 病(HMSN1)	普遍的分子机制：非线性重复序列的重组	17p11.2 的 1.5Mb 重复(图 16.2)；也发生点突变
α 和 β 地中海贫血	杂合子选择导致不同的始祖突变，在不同的群体中常见不同的始祖突变	见图 16.2(α 地中海贫血)和表 18.5(β 地中海贫血)
泰-萨(Tay-Sachs)病	在 Ashkenazi 犹太人有建立者效应，远古杂合子优势	在 Ashkenazim 犹太人的两种常见 *HEXA* 基因突变：在外显子 11 有 4bp 插入(73%)；外显子 11 供体剪接位点 G→C(15%)
囊性纤维化	在北欧群体中常见的始祖突变，远古杂合子优势	见表 18.6 及节 4.5.3

- 家系内诊断。可能需要突变筛查方法来确定家系突变，但是一旦对突变定性，家系中其他成员通常仅需检测这种特定的突变即可；
- 研究中，用于检测对照样本。定位克隆的一个常见问题是患者的候选基因有序列改变，这就引出了问题，这个改变是致病性的（证明候选基因为致病基因），还是一种非致病性的多态性？一个常用的方法是筛查一组正常的对照样本，确定是否存在此种改变。Collins 和 Schwartz (2002) 讨论了多少对照用于检测的问题；
- SNP 基因分型。在这里，目的不是如上面那样寻找致病性突变。但问题是相同的：都是检测 DNA 样品的一种已预先确定的序列变异。SNP 分型是超高通量基因分型的主要应用（节 18.4.2）。

18.4.1　许多简便的方法可用于特定变异体基因分型

检测限制性位点的存在或缺如

当某一碱基替换产生一个识别位点或使限制酶识别位点消失时，以单纯的直接 PCR 方法即可检测突变（表 7.6）。虽然已知有数百种限制酶，但他们都识别对称的回文结构位点。并且许多点突变不会影响这样的序列。此外，罕见的和不明确的限制酶不适合于常规诊断使用，因为这些酶价格昂贵，质量通常较差。然而有时可以通过使用仔细选择引物的 PCR 诱变作用（节 5.5.3），引入一个具诊断意义的限制位点。图 18.8 给出了一个例子。

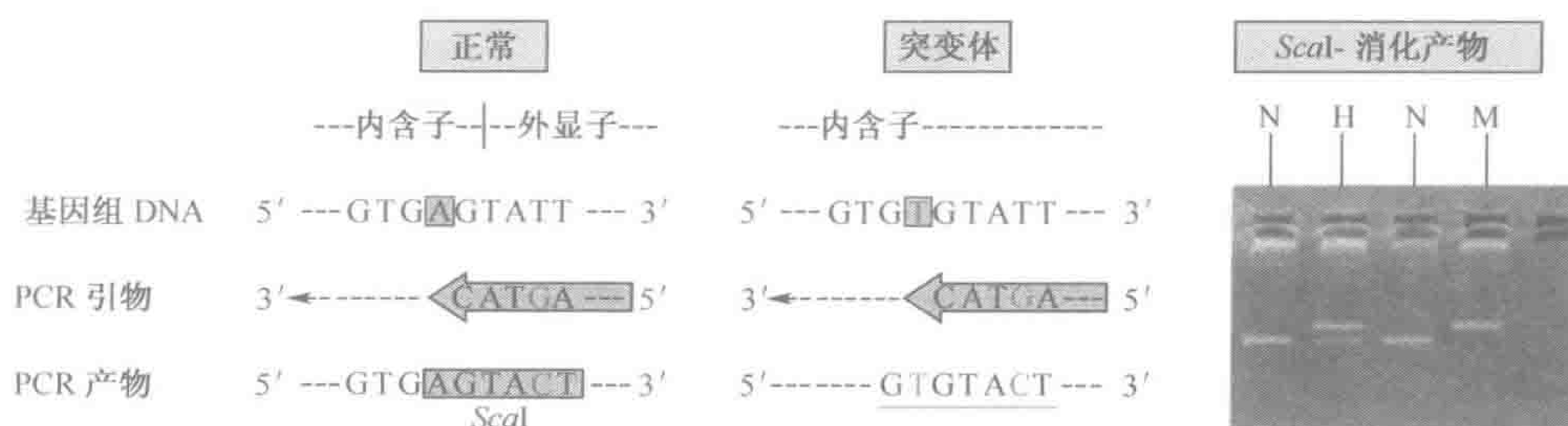


图 18.8　引入人工诊断性限制位点

FACC 基因内含子 4 剪接位点 A→T 突变并不产生，也不使限制性酶切位点消失。PCR 引物止于该变化的碱基前，但在非关键位置有单碱基错配（灰色 G），这并不妨碍与正常序列和突变序列的杂交和扩增。引物错配使正常序列的 PCR 产物引入 *Sca*I 的 AGTACT 限制性酶切位点。图中所示为纯合子正常（N），杂合子（H），纯合子突变（M）经 *Sca*I 消化的产物条带。英国伦敦 Guy 医院 Rachel Gibson 博士惠赠。

等位基因特异性寡核苷酸（ASO）杂交的应用

在适度严格杂交的条件下，这些短的合成探针只能与完全匹配的序列杂交（节 6.3.1）。图 6.11 阐明了应用 ASO 探针进行斑点杂交检测引起镰状细胞病的单碱基替换。出于诊断目的，经常使用反斑点杂交程序，例如以筛查一系列已明确的囊性纤维化突变为例，应该采用一系列与每种突变等位基因特异的 ASO 探针，将它们点样到单层膜上，随后与标记的经 PCR 扩增的待测 DNA 杂交。

相同的反斑点杂交原理被应用于 DNA 芯片大规模检测。成千上万的 20～25 个碱基的寡核苷酸探针锚定在固体支持物的特定位置（节 6.4.3）。待测 DNA 经 PCR 扩增、荧光标记后与微阵列进行杂交，或单独杂交，或（更为可取的是）与标记了不同颜色的野生型参照序列进行竞争。芯片中每一个单独的探针完成一个识别待测 DNA 上特定序列的 ASO 检测，但总体上，微阵列能够筛查某一基因的全部序列，并能检出任何碱基替换（图 18.9A）。对纯合子的检测效率非常高，对杂合子其可靠性要差一些。只要预先将探针设计成检测特定的插入突变，该检测方法才能检出插入。

尽管比较昂贵，一旦每件事情都准备完善，芯片系统就能快速而方便地被使用。然而，有必要预先明确待检突变的范围，并将这些设计入芯片中。因此芯片系统在突变检测中的主要作用就是发挥初始筛查样品的作用，检出常见突变，把一些棘手的情况留给

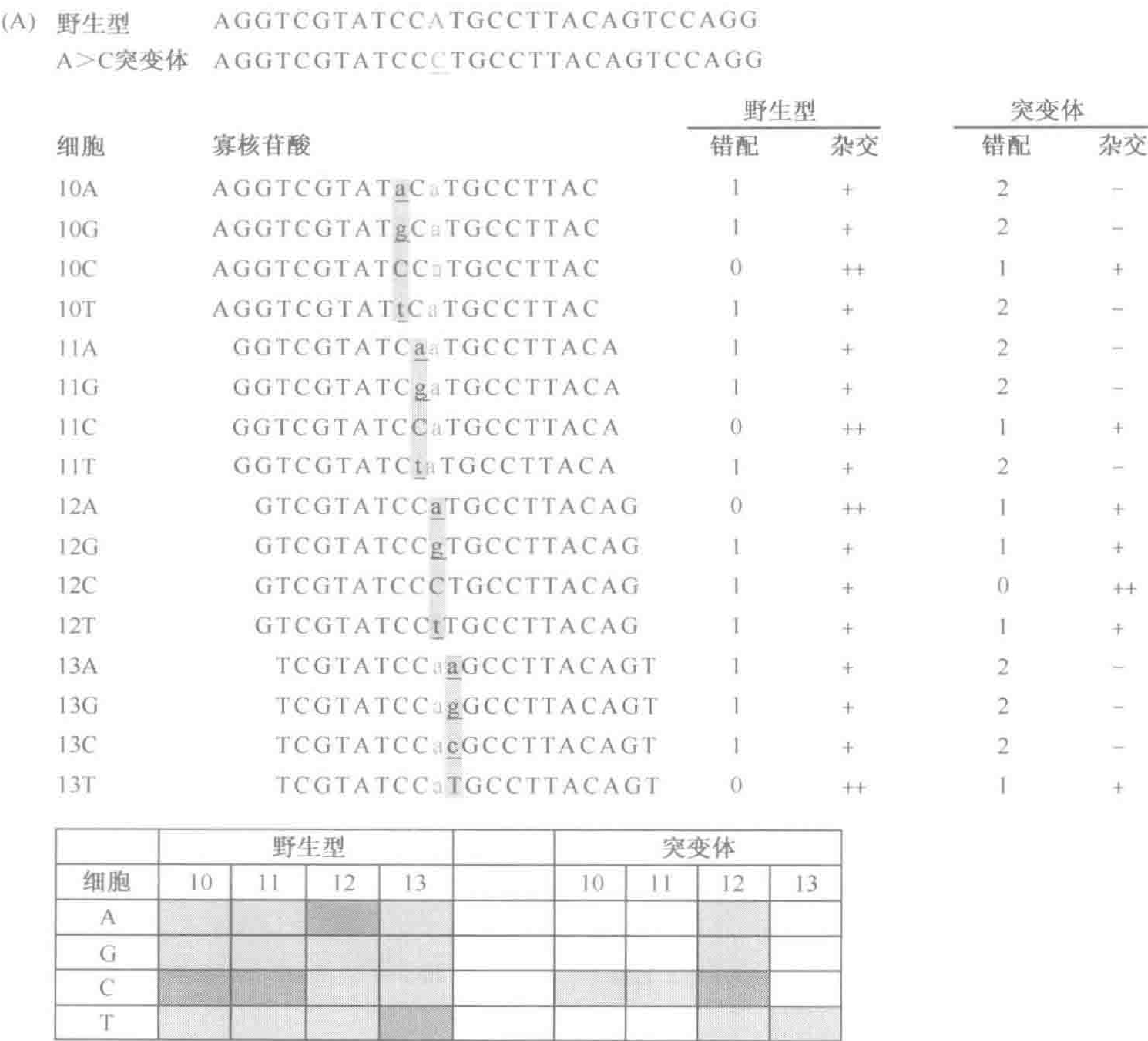


图 18.9　寡核苷酸阵列用于突变检测

(A) 杂交检测突变的原理。寡核苷酸排列成四组，每组对应某一位置（阴影）四种可能的碱基。与野生型序列的错配以小写字母表示，正确配对以阴影中大写字母表示。突变序列在位置 12 处有 A→C 替换。当野生型或突变序列与阵列杂交时，错配的数目和杂交的强度显示于右侧。该表显示了这一过程。图 7.4 有一个实例。(B) 小测序阵列的原理。阵列的每个小室含有一个与部分靶序列相匹配的寡核苷酸。寡核苷酸锚定于其 5′端，与靶序列杂交后，寡核苷酸成为单碱基延伸的引物，该过程采用彩色标记的双脱氧 NTP。3′端的错配阻止延伸；在尾部错配一个或两个碱基延伸反应较弱。

(B)		
野生型	CTAGTTCGACGAGGTCGTATCCATGCCTTACAGTCCAGG	
靶 DNA	CTAGTTCGACGAGGTCGTATCCCTGCCTTACAGTCCAGG	
		添加的核苷酸
引物 1	5′ CTAGTTCGACGAGGTCGTA 3′	ddT- 红色
引物 2	5′ TAGTTCGACGAGGTCGTAT 3′	ddC- 蓝色
引物 3	5′ AGTTCGACGAGGTCGTATC 3′	ddC- 蓝色
引物 4	5′ GTTCGACGAGGTCGTATCC 3′	ddC- 蓝色
引物 5	5′ TTCGACGAGGTCGTATCCA 3′	未加标签
引物 6	5′ TCGACGAGGTCGTATCCAT 3′	ddG- 黄色（非常弱）
引物 7	5′ CGACGAGGTCGTATCCATG 3′	ddC- 蓝色（弱）
引物 8	5′ GACGAGGTCGTATCCATGC 3′	ddC- 蓝色
引物 9	5′ ACGAGGTCGTATCCATGCC 3′	ddT- 红色
引物 10	5′ CGAGGTCGTATCCATGCCT 3′	ddT- 红色
引物 11	5′ GAGGTCGTATCCATGCCTT 3′	ddA- 绿色

图 18.9（续）

其他方法解决。

等位基因特异性 PCR 扩增（ARMS 检测）

ARMS（扩增阻碍突变系统）的原理见图 5.4。进行配对 PCR 反应，一个引物（通用引物）在两个反应中相同，而另一个引物在两个反应中稍有不同。一个是正常序列特异引物，另一个是突变序列特异引物。此外通常包括对照引物，用于扩增每个样品中一些无关序列以检验 PCR 反应是否正确。可以选择通用引物结合位点，以针对不同的突变，产生不同长度的产物。因此，多重反应 PCR 的产物在凝胶上形成梯状。在仔细设计引物情况下，突变特异的引物也可以产生有差别的产物。例如，他们可以被不同的荧光素或其他标记物标记，或产生不同长度的 5′延伸产物。多重突变特异 PCR 非常适合筛查相对大量样本的一组特定突变（图 18.10）。

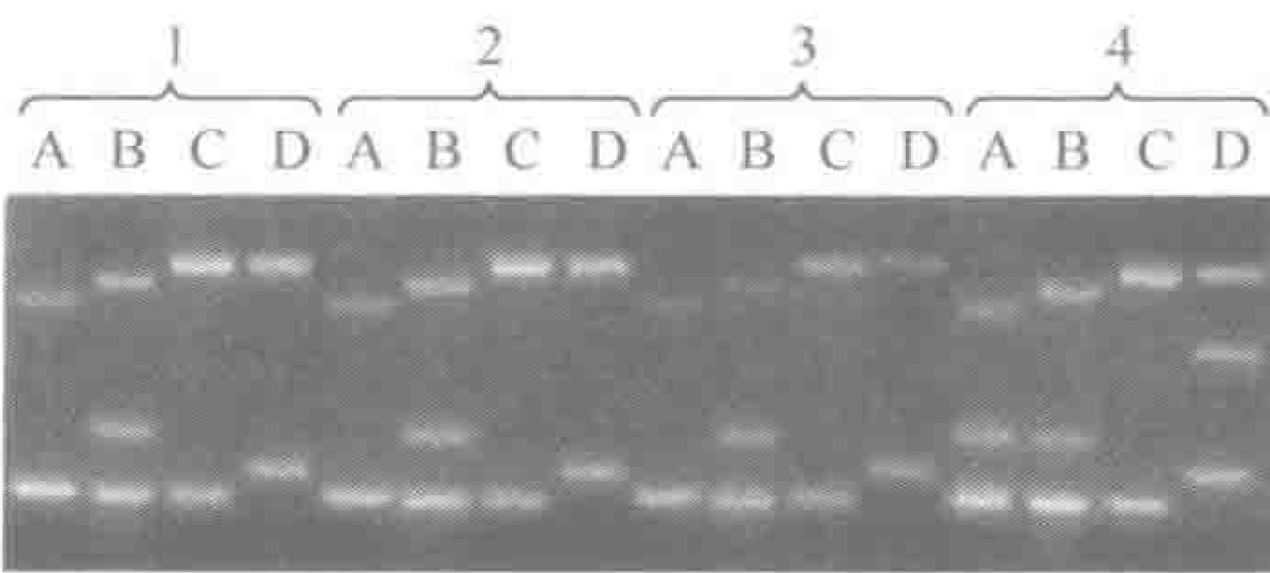


图 18.10　多重 ARMS 实验检测 29 种囊性纤维化突变

用四个多重突变特异 PCR 反应（A-D）来检测每一个样品。在一个多重反应中，每个产物大小不同。如果存在 29 种突变中的一种，就会出现多余的一条带。通过判断凝胶中带的位置及其所在泳道，就能识别出突变。每个管中也扩增两个对照序列，他们是每个凝胶泳道的上部和下部条带，在不同泳道之间有差别，因此每个多重反应都具有自己的标记模式。注意：除了 F508del 突变（条带在 B 泳道），其他突变的正常等位基因未被检测。样品 1，2，3 没有突变特异条带。对于样品 4，泳道 A 和 D 中额外的条带分别代表 F508del 和 1898+1G→A，表明该 DNA 来源于复合杂合子患者。英国 Manchester St Mary 医院 Michelle Coleman 博士惠赠。数据采用 Orchid Biosciences 生产的 Elucigene™试剂盒获得。

寡核苷酸连接分析（OLA）

OLA 检测碱基替换突变时，所构建的两个寡核苷酸与靶基因的邻近序列杂交，连接点位于突变位的位置。除非两个寡核苷酸能够完全杂交（Nickerson *et al.*，1990），否则 DNA 连接酶不能将二者共价相连。多种检测方式是可能的，如 ELISA，或图 18.11 中荧光测序仪进行分析。

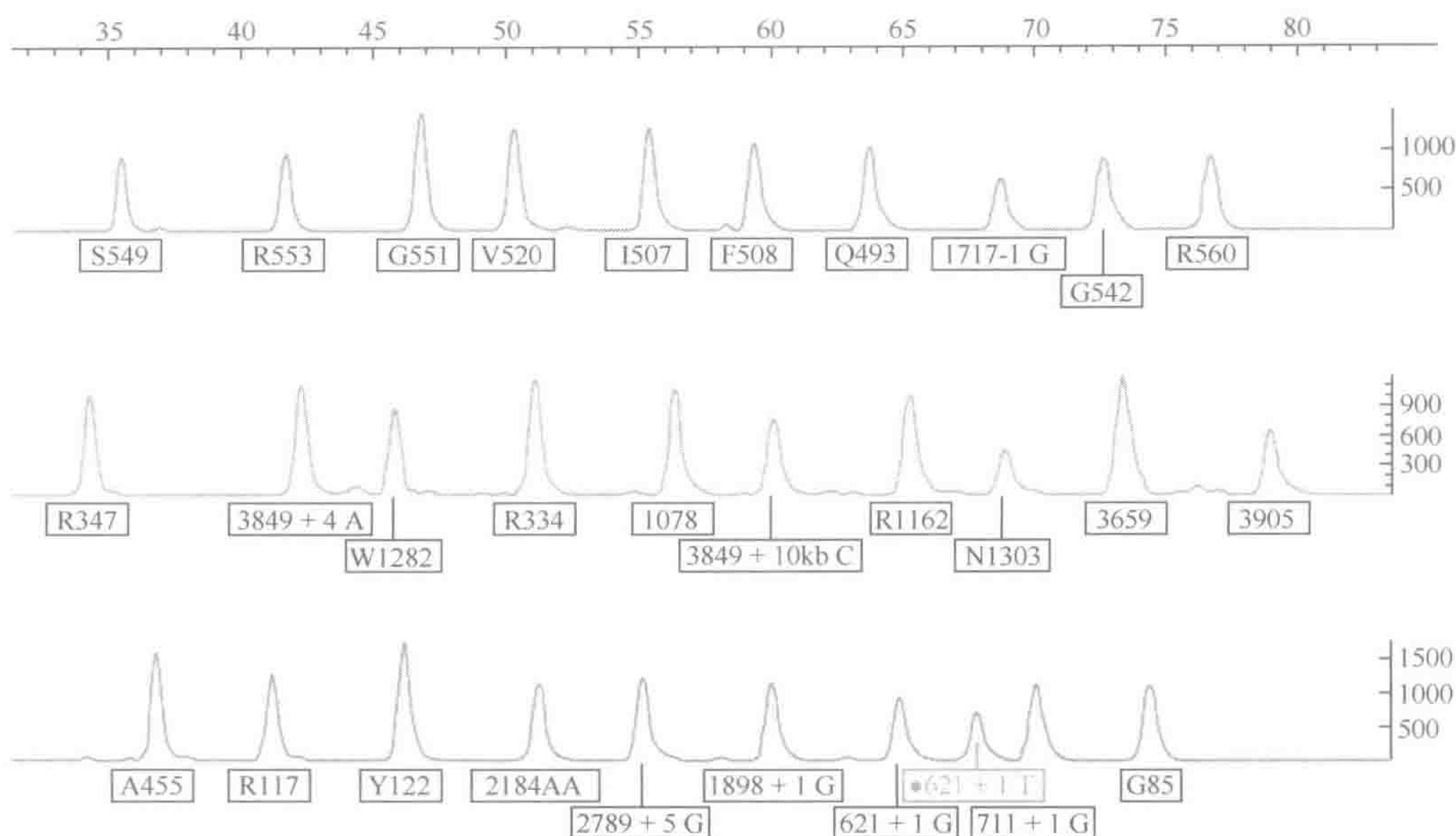


图 18.11 用寡核苷酸连接分析检测 31 种已知的囊性纤维化突变

多重 PCR 后，进行多重寡核苷酸连接分析。对连接寡核苷酸加以设计，使每个突变和其正常对照产物能通过大小和标记的颜色加以区别，可以见到剪接位点突变 621＋1g→t 的连接产物。受检者可能是携带者或是携带第二种突变的复合杂合子，这种突变不在该试剂盒检测的 31 种突变中。英国 Manchester St Mary 医院 Andrew Wallace 博士惠赠。数据采用 ABI Biosystems 试剂盒获得。

通过引物延伸进行小测序

小测序利用 Sanger 测序（图 7.2）原理，但仅加入一种核苷酸。测序引物的 3′端位于待确定基因分型核苷酸的直接上游区，反应混合物含有 DNA 多聚酶加上四种标记各异的双脱氧三磷酸核苷（ddNTP），以待测 DNA 为模板向引物上添加一种已标记的双脱氧核苷酸，采用这样或其他的方法，添加到引物上的核苷酸得以识别（例如，通过在荧光测序仪上对延伸的引物测序），从而识别目的位置的碱基。

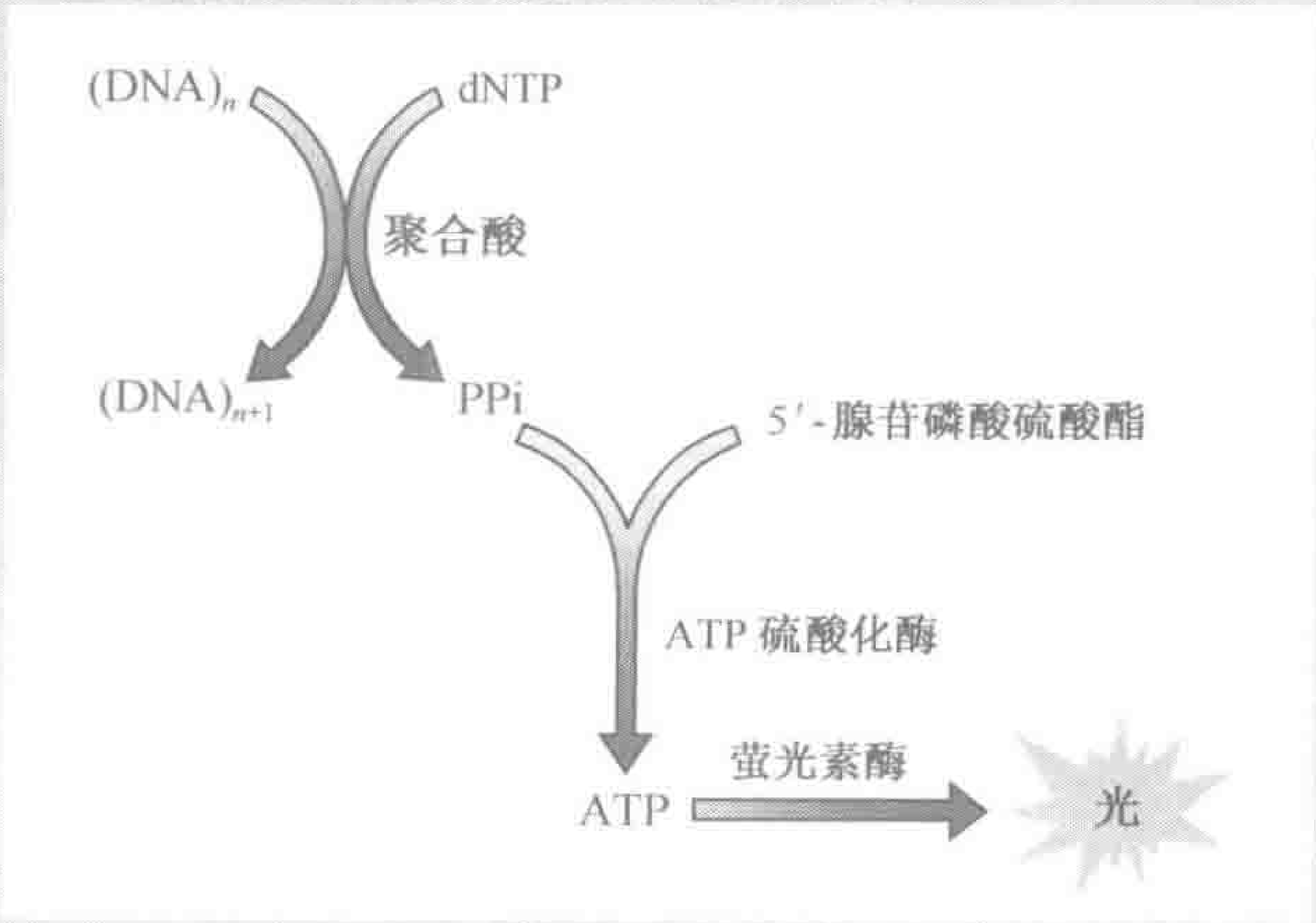
小测序易于转换成适合阵列型研究的方式（APEX，arrayed primer extension；Tõnisson *et al.*，2002），通过一次操作完成待检基因整个序列碱基替换的检测（图 18.9B）。

18.4.2　高通量基因分型方法

生物技术公司正在开发无数的系统，以满足用于疾病相关研究的巨量 SNP 基因分型的挑战（节 15.4.5），Weaver（2000）列出了大量的这类系统。根本的遗传学方法通常是如上所述的引物延伸（小测序）、等位基因特异寡核苷酸杂交或寡核苷酸连接，但已实现自动化和超高通量。许多系统尝试将遗传学技术和实验室芯片开发用微液流技术结合起来。其他方法包括焦磷酸酸测序和质谱测定法（框 18.2）。

框 18.2　两种高通量基因分型方法

焦磷酸测序：这是一种检测紧邻某确定起始点超短片段序列的方法。主要用途是仅有 1～2 个碱基需要测序的 SNP 分型。当一个 dNTP 掺入到新生 DNA 链时，会有焦磷酸释放出来，焦磷酸测序采用一组巧妙组合的酶偶联释放出的焦磷酸，由荧光素酶发出光信号（Fakhrai-Rad *et al.*，2002）。采用该方法的仪器每天能实现自动化分析 10 000 个样品。输出结果是定量的，因此 SNP 的等位基因频率可以在大量混合样品中的一次分析中估计出来。



MALDI-TOF MS（基质辅助激光解吸附/离子化时间飞行质谱）在真空中将离子加速到达靶目标，通过计算飞行时间或者测量离子在磁场中偏离的距离，质谱测量质量：离子电荷比（详见框 19.7）。采用质谱分析诸如 DNA 或蛋白质等大分子的问题是获得自由飞行的离子。MALDI 技术通过把这个大分子包埋到一个吸光物质的微点上解决这一问题，该微点随后被短暂的激光脉冲所蒸发（Monforte and Becker，1997）。飞行到靶目标的时间与质量：电荷比的平方根成正比。

应用到 DNA，该技术能检测大到 20kDa 质量的物质，准确度为±0.3%。质谱能作为一种非常快的方法，代替凝胶电泳检测长度大到大约 100 个核苷酸的寡核苷酸。典型的应用包括分析 Sanger 测序反应产物或测定微卫星大小。对于小的寡核苷酸，其准确度足以通过其确切质量直接推断出碱基组成。另一方面，采用质量标记 ddNTP，通过引物延伸可以对 SNP 进行分析。将待分析 DNA 的样品排列于盘上，机器自动依次把每个点离子化。目前的体系每天可以对成千上万 SNP 进行基因分型。该技术与焦磷酸测序结合时，样品可混合在一起，直接测定等位基因频率。

18.4.3　三核苷酸重复疾病的遗传检测

导致许多神经性疾病（表 16.6）的扩张重复涉及一套特殊的突变特异性检测（图

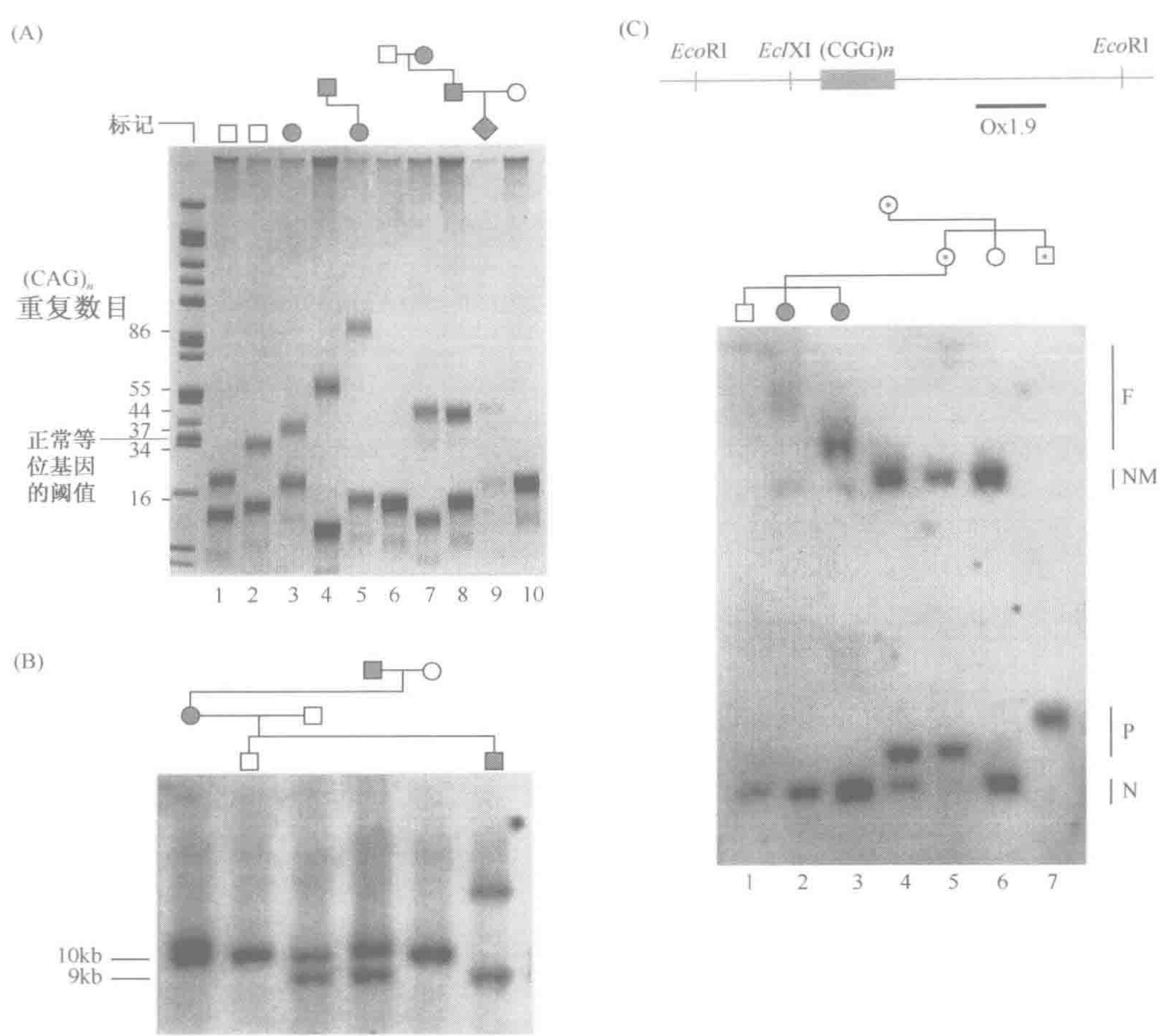


图 18.12 三核苷重复疾病的实验室诊断

(A) 亨廷顿病：含有 $(CAG)_n$ 重复的基因片段经 PCR 扩增和聚丙烯酰胺凝胶电泳分离，条带经银染显示。标尺显示了重复数目。泳道 1，2，6 和 10 来自未受累者样品，泳道 3，4，5，7 和 8 来自受累者样品。泳道 5 是青春期发病的病例；他父亲（泳道 4）有 45 个重复片段，但她有 86 个重复片段，泳道 9 是产前诊断的受累胎儿。英国 Manchester St Mary 医院 Alan Dodge 博士惠赠。(B) 肌营养不良：经 *Eco* RI 消化的 DNA Southern 印迹，9 或 10kb（箭头）条带是正常突变体。祖父有白内障但没有其他肌营养不良症状，他的 10kb 条带似乎是非常轻微地扩张，但仅以这块凝胶为证据，扩张还是不明确的。他的女儿有一条正常和一条确定无疑扩张的 10kb 条带。她有典型的成人期发作性肌营养不良，她的儿子有一个大的扩张条带，并有严重的先天性疾病。英国 Manchester St Mary 医院 Simon Ramsden 博士惠赠。(C) 脆性 X 染色体病。女性失活的 X 染色体上的 DNA 以及任何携带有全突变的 X 染色体上的 DNA 被甲基化。用 *Eco*RI 和甲基化敏感酶 *Ecl*XI 联合消化 DNA，Southern 印迹并与 0×1.9 或相似的探针杂交。与正常男性 X 染色体（泳道 1）和女性有活性正常 X 染色体（泳道 2，3，4，6）杂交后产生的片段小（标记为 N)。非甲基化前突变等位基因（P）在泳道 4 和 5（女性前突变携带者）和泳道 7（正常传递疾病的男性）有一个稍大一点的条带。甲基化（失活）X 染色体序列不能被 *Ecl*XI 切割，产生更大的条带（NM），而全扩张和甲基化序列因为体细胞嵌合性有非常大的弥散条带 (F)。英国 Manchester St Mary 医院 Simon Ramsden 博士惠赠。

18.12)。对于多聚谷氨酰胺重复性疾病，如亨廷顿病（HD），单一的 PCR 反应可作出

诊断。其他一些扩张性重复疾病诊断较困难，原因有两点。全突变可能包含成几百或几千个重复序列，不易经 PCR 扩增，特别是大多数重复具有很高的 GC 含量。正常和前突变的等位基因可产生清晰的 PCR 产物，但全突变可能需要 Southern 印迹。此外，与 Huntington 病不同，突变大多数通过功能丢失导致疾病，偶发的受累患者有缺失和点突变，如果仅检测重复，将会被漏检。肌营养不良是“大型扩张性”疾病中唯一一个似乎是完全纯合突变的疾病。

18.4.4 对于一些检测，地理来源是一个重要的考虑因素

隐性遗传病的群体遗传学常受到建立者效应或杂合子优势效应的支配，所导致的群体中突变的有限多样性使遗传检测变得更加容易。β 地中海贫血和囊性纤维化就是很好的例子。对于这两种疾病，已报道了相关基因的大量不同类型的突变，但在任何特定的群体中，每种疾病中少数的突变是大多数病例的原因。对于 β 地中海贫血，不需要通过 DNA 检测来诊断携带者或受累者，传统血液学可准确诊断，但 DNA 检测可作为产前诊断的一个选择。在不同的群体中，不同的突变占有绝对优势（表 18.5）。假如可以从双亲处获得 DNA 样本，并知道他们的种族来源，就可以通过几种特异性检测来找出亲代的突变，这样胎儿很容易作出检测。

表 18.5　在不同的国家中主要的 β 地中海贫血突变

在每一个国家中，由于建立者效应和以及合子选择优势联合作用，某些突变是常见的。

群体	突变	频率(%)	临床效应
撒丁岛	密码子 39(C→T)	95.7	β^0
	密码子 6(delA)	2.1	β^0
	密码子 76(delC)	0.7	β^0
	内含子 1～100(G→A)	0.5	β^+
	内含子 2～745(C→G)	0.4	β^+
希腊	内含子 1～100(G→A)	43.7	β^+
	密码子 39(C→T)	17.4	β^0
	内含子 1～1(G→A)	13.6	β^0
	内含子 1～6(T→C)	7.4	β^+
	内含子 2～745(C→G)	7.1	β^+
中国	密码子 41/42(delTCTT)	38.6	β^0
	内含子 2～645(C→T)	15.7	β^0
	密码子 71/72(insA)	12.4	β^0
	−28(A→G)	11.6	β^+
	密码子 17(A→T)	10.5	β^0
巴基斯坦	密码子 8/9(insG)	28.9	β^0
	内含子 1～5(G→C)	26.4	β^+

续表

群体	突变	频率(%)	临床效应
美国非洲黑人	619bp 缺失	23.3	β^+
	内含子 1～1(G→T)	8.2	β^0
	密码子 41/42(delTCTT)	7.9	β^0
	－29(A→G)	60.3	β^+
	－88(C→T)	21.4	β^+
	密码子 24(T→A)	7.9	β^+
	密码子 6(delA)	0.8	β^0

英国牛津 Weatherall Molecular 医学院 J. Old 博士惠赠资料。

在囊性纤维化中，F508del 突变在所有的欧洲群体中最常见，被认为起源于远古。然而，在所有突变中，F508del 所占的比例不同，通常在北欧和西欧高而南部较低。囊性纤维化突变检测分为两个阶段，首先，少数的特异突变，总是要包含 F508del 突变，可用 18.4 节介绍的方法来寻找。正如表 18.6 所示，以检测特定突变效果回报递减而言，并没有明显的自然临界值。如果这个阶段没能检测出突变，那么如果条件允许，可依照节 18.3 叙述的或基因示踪的方法（图 18.13）进行未知突变的筛查。下面讨论了这种多样性对群体筛查计划的影响（图 18.18）。

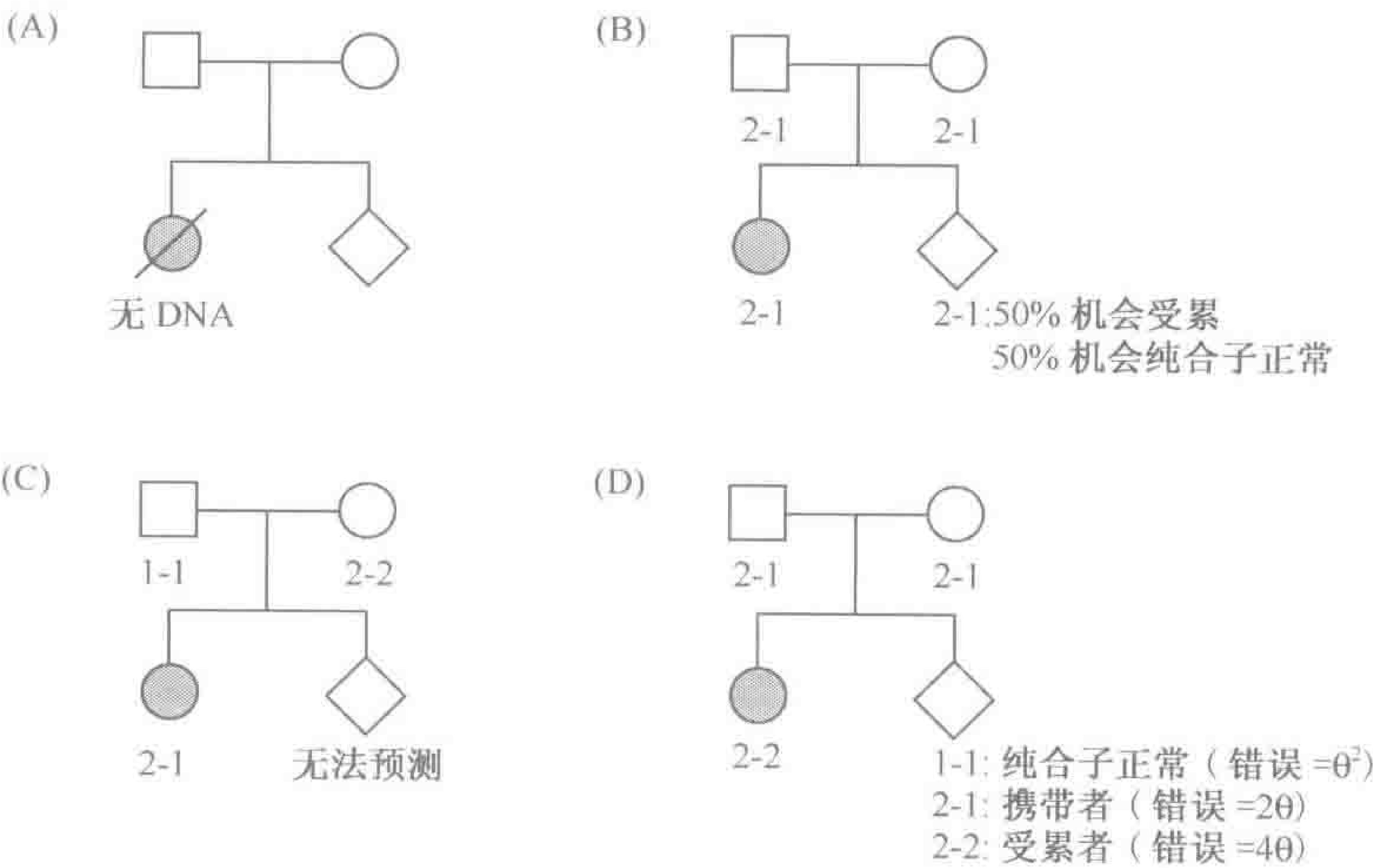


图 18.13 常染色体隐性疾病产前诊断的基因示踪

（A）没有 DNA（B）50%机会受累 50%机会纯合子正常（C）无法预测（D）纯合子正常（错误＝θ^2）携带者（错误＝2θ）受累者（错误＝4θ）

四个家庭，每个家庭中有一个患隐性疾病的小孩，直接突变检测是不可能的（可能是因为该基因尚未被克隆，或是因为突变都不能被找到）。（A）如果无受累儿童样品，则无法诊断。（B）如果每个人对于标记物均是相同的杂合子基因型，这样结果没有临床意义。（C）双亲对于标记物均是纯合子，也无法利用这个标记物进行预测。（D）成功的预测。如果所使用的标记物与致病基因座的重组率为 θ，图中显示的错误率是指当胎儿实际受累但预测为未受累妊娠的风险，或指当胎儿实际未受累但预测为受累妊娠的风险。这些例子强调了适当的系谱结构（受累儿童的 DNA 样品必须可以获得）和能提供信息量的标记类型的必要性。

当一特定群体中一种隐性遗传病特别常见时，结果通常是一个以上的突变是疾病的原因。例如，Ashkenazi 犹太人中的泰-萨病，就存在两个常见的 *HEXA* 突变（表 18.4）。难以解释这一现象，除非假定创立的群体较小时，存在一个较大的杂合子优势。

表 18.6　来自英格兰西北部 300 例囊性纤维化染色体中 CFTR 突变的分布

突变	外显子	频率(%)	累计频率(%)
F508del	10	79.9	79.9
G551D	11	2.6	82.5
G542X	11	1.5	84.0
G85E	3	1.5	85.5
N1303K	21	1.2	86.7
621+1 G→T	4	0.9	87.6
1898+1 G→A	12	0.9	88.5
W1282X	21	0.9	89.4
Q493X	10	0.6	90.0
1154insTC	7	0.6	90.6
3849+10kb(C→T)	intron19	0.6	91.2
R553X	10	0.3	91.5
V520F	10	0.3	91.8
R117H	4	0.3	92.1
R1283M	20	0.3	92.4
R347P	7	0.3	92.7
E60X	3	0.3	93.0
未知/隐蔽的突变	—	7.0	100

F508del 和其他几种相对常见的突变可能是来自古代并通过杂合子选择优势传递下来。其他的突变可能是近代的，罕见并具有高度异质性。囊性纤维化在这类群体中比大多数其他群体更具有同质性。突变的命名法见框 16.2。资料由英国 Manchester St Mary 医院 Andrew Wallace 博士惠赠。

18.5　基因示踪

基因示踪是历史上第一种广为使用的 DNA 诊断方法。它使用的是疾病基因座定位方面的知识，而不是实际疾病基因方面的知识。因此，对已被定位的基因，而不是已克隆的基因，它是唯一可行的方法。大部分构成诊断性实验室日常工作的孟德尔式疾病经历基因示踪的阶段，一旦基因被克隆，就转向直接的检测。亨廷顿病、囊性纤维化和肌营养不良就是熟知的例子。然而，即使当基因已被克隆，基因示踪仍可能发挥作用。在诊断性实验室条件下，对一个大的多外显子基因进行全序列探查每个突变并不总是经济的。此外，18.3 中介绍的突变筛查方法中没有一个灵敏度可达到 100%。因此，总会有突变不能被发现的病例。在这些情况下，使用连锁标记物的基因示踪是一个可供选择的方法。基因示踪的必要条件是：

1. 疾病应当被充分定位，这样才可以使用已知的与疾病基因座紧密连锁的标记物。

2. 系谱结构和样品来源必须允许确定位相（phase）（见下文）。

3. 必须有明确的临床诊断，对疾病基因的定位不能存在不确定性。

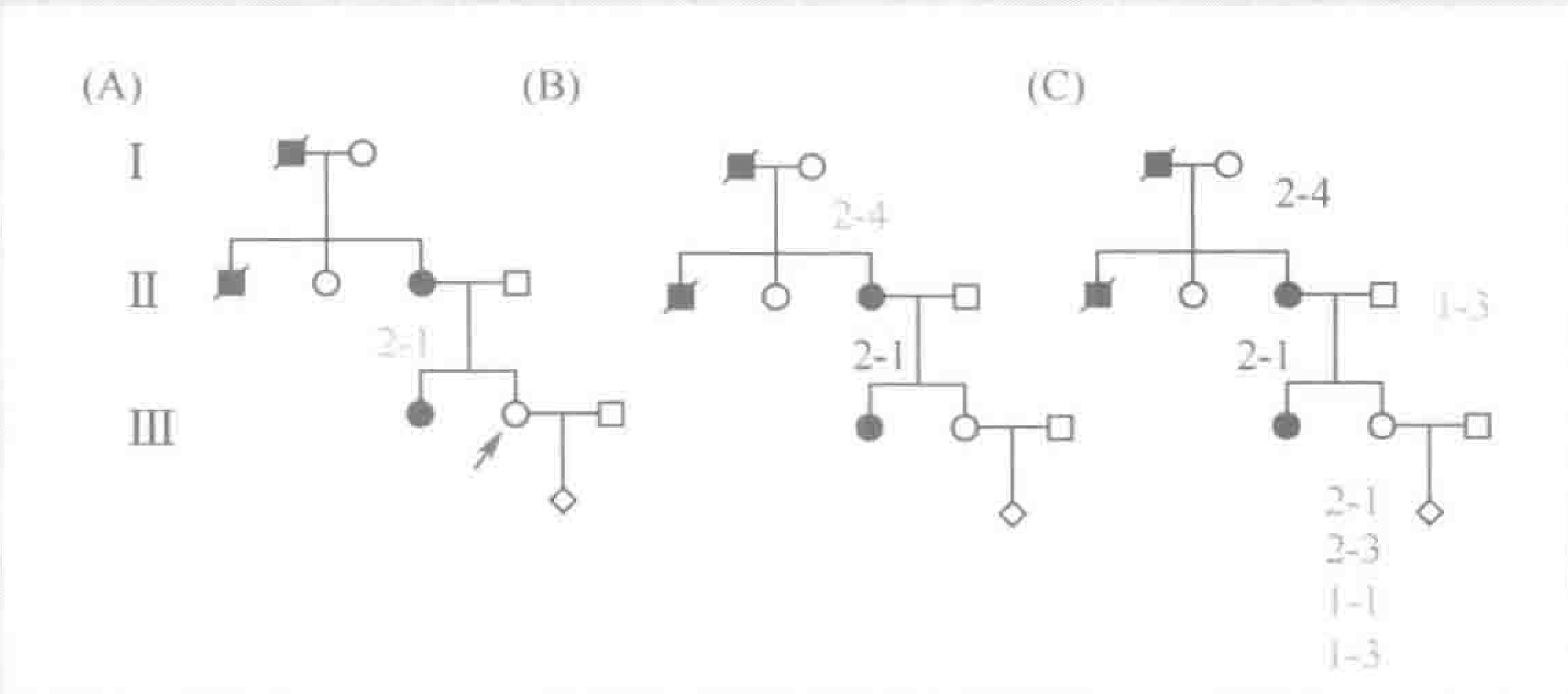
18.5.1 基因示踪包括三个逻辑性步骤：

框 18.3 说明了基因示踪的基本逻辑，它适用于任何遗传类型的疾病。总是至少有一位亲代将疾病等位基因传给先证者，谁在事实上可能或不可能传递了疾病等位基因。该过程总是遵循相同的三个步骤：

1. 区分相关亲代中的两条染色体，即找到对他们来说是杂合子的紧密连锁的标记。
2. 确定位相（phase）：即找到哪条染色体携带疾病等位基因。
3. 找到咨询者遗传到的是哪条染色体。

框 18.3 基因示踪的逻辑

迟发性常染色体显性疾病调查的三个步骤，由于这样或那样的原因，直接检测这种疾病突变是不可能的。



(A) Ⅲ-2（箭头），已妊娠，欲作症状前检测以确认她是否已经遗传了致病等位基因。第一步是区分她母亲的两条染色体。已找到一个与疾病基因座紧密连锁的标记，对此，Ⅱ-3 是杂合子。

(B) 下一步，我们必须建立位相，也就是要判断Ⅱ-3 中的哪个标记等位基因与疾病等位基因分离。Ⅰ-2 以该等位基因进行分型，Ⅱ-3 一定已经从她母亲那里遗传了标记等位基因 2，因此可用于标记她的未受累染色体，从她死亡父亲那里遗传的受累染色体一定是携带标记等位基因 1 的那一条。

(C) 通过对Ⅲ-2 和她父亲基因型分析，我们可以判断她从母亲那里获得了哪一个标记等位基因，如果她是 2-1 或 2-3，这是个好消息：她从她母亲那里遗传了标记等位基因 2，这是她祖母的等位基因。如果她的基因型为 1-1 或 1-3，这是个坏消息：她遗传了她祖父的染色体，该染色体携带了疾病等位基因。

注：这是家系中的分离模式，并不是实际的标记基因型，这一点是重要的：如果Ⅲ-2 有与其受累母亲同样的标记基因型 2-1，这对她是个好消息，不是坏消息。

18.5.2 重组从根本上限制了基因示踪的准确性

因为用于基因示踪的 DNA 标记不是导致疾病的序列，如果重组将疾病与标记分

离，总会存在预测错误的可能性。重组部分，也就是错误率，可以通过标准的连锁分析（章 13）从家系分析中估计出来。对于几乎任何疾病，应该有很多与疾病基因座重组率小于 1%的标记供选择。这是根据以下观察推出的结论：每 300 个核苷酸中就有一个是多态的，相距 1Mb 的基因座间重组率大约为 1%（节 13.15）。最理想的是采用基因内的标记，例如，内含子中的微卫星。下面将提到杜兴肌营养不良重组热点的特殊问题。

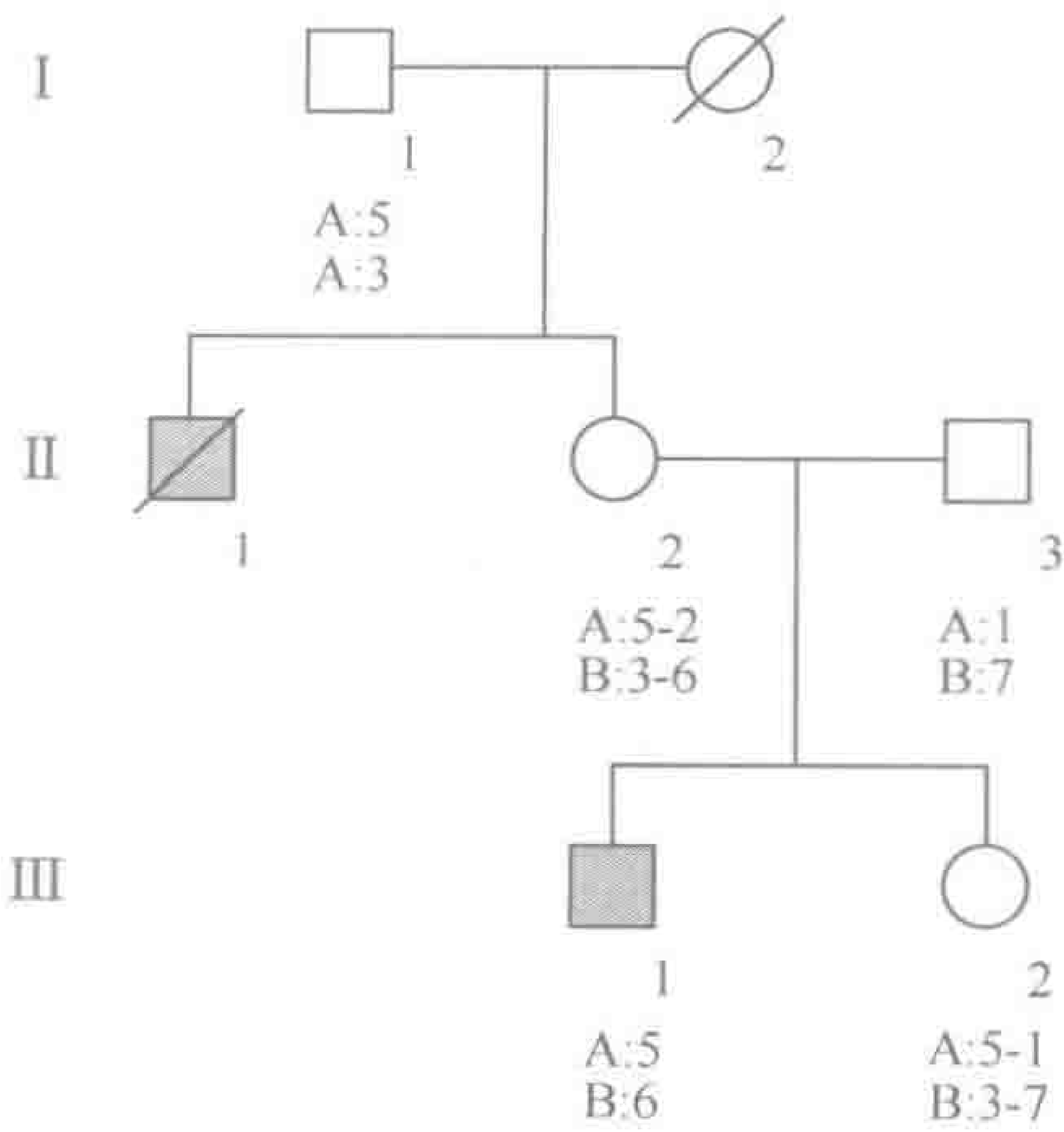


图 18.14　使用旁侧标记对杜兴肌营养不良的基因示踪

利用抗肌萎缩蛋白基因座侧翼的两个多态位点 A 和 B 对该家系进行分型，只有在标记 A 和 DMD 有一次重组以及 DMD 和标记 B 之间有另一次重组，Ⅲ-2 才遗传了 DMD。如果重组率分别为 θ_A 和 θ_B，双重重组体的概率是 $\theta_A \times \theta_B$，这一概率将远低于 1%。Ⅲ-1 在标记基因座 A 和 DMD 之间有一次重组。

即使对非常紧密连锁的标记，也不能完全排除标记与疾病之间的重组。但是利用位于疾病基因座两侧的两个标记可以极大地减少错误率，通过这种**旁侧**（flanking）或称**桥联的标记**（bridging marker），在疾病基因座与其中之一的标记产生重组时也会形成标记-标记重组体，这可以被检测到（例如图 18.14 中 Ⅲ-1）。如果标记-标记重组体在咨询者中被发现，那么就不能对疾病的遗传做出预测，但是至少避免了错误的预测。假如没有发现标记-标记重组体，唯一剩余的风险就是双重重组体。由于干扰（节 13.1.3），双重重组体的真实的可能性是非常低的。这样，由于未注意到的重组产生的错误率远远比在 DNA 样品获得和处理中人为错误所导致的错误预测率低得多。也许，更大的风险是未预见到的基因座杂合性，以至于家系中真正的疾病基因座与正在被跟踪的基因座实际上是不同的。

18.5.3　计算基因示踪中的风险

与直接突变检测不同，基因示踪总是涉及到计算。评价最终风险需考虑的因素包括：

- ▶ 疾病-标记和标记-标记重组的可能性；
- ▶ 不确定性：关于谁传递了什么标记等位基因给谁，这是由于不完整的系谱结构或标记有限的信息含量（见图 13.5C 中的例子）；
- ▶ 对于在这个家系中是否有人携带了新的突变的疾病等位基因，存在不确定性（见图 4.8 DMD 中该问题的一个例子）。

对此进行计算，有两种可供选择的方法。

Bayes 计算（框 18.4）

Bayes 定理提供了将概率组合成为一个最终总体概率的一般方法。其原理和方法见框 18.4。图 18.15 中有一个示例计算。在 Bridge（进一步阅读）所著的书中可以找到

一套非常详细的、涵盖了在DNA诊断中的几乎所有可能情况的计算方法，感兴趣的读者可以查阅此书。

对于简单的系谱，Bayes计算可以快速给出答案。而对于比较复杂的系谱，该计算就变得比较繁琐。尽管试图揭示决定最终风险的因素是一项有价值的脑力劳动，很少有人对自己能正确分析复杂系谱的能力信心十足。另一种方法是采用连锁分析程序。

框 18.4 Bayes定理在联合概率中的应用

Bayes原理的公式是：

$$P(Hi \mid E) = P(Hi).\ P(E \mid Hi)/\Sigma[P(Hi).\ P(E \mid Hi)]$$

P(Hi) 意思是 i^{th} 假设的概率，竖线表示“在... 条件下”，所以P（E | Hi）表示在假设Hi的条件下，证据（E）的概率。举一个例子可能会使其更加清晰，Bayes计算的步骤是：

Ⅰ为每种假设建立一个一列的表格，把所有的可能都包括进去。

Ⅱ为每种假设分配一个**前概率**（prior probability），所有假设的前概率之和必须等于1。在这个阶段你担心应该用怎样的信息来决定前概率是不重要的，只要信息在每列间保持一致即可。你将不会用到所有的信息（否则进行计算是没有意义的，因为你已经有了答案），在前概率中没有被用到的信息在后来能被用到。

Ⅲ用一条在前概率中未被包括在内的信息，为每种假设计算**条件概率**（conditional probability）。条件概率是在假设的条件下，即P（E | Hi）的条件下，信息的概率［不是在信息P（Hi | E）的条件下，假设的概率］。不同假设的条件概率总和不一定等于1。

Ⅳ如果还有没有包括在内的进一步信息，重复步骤（Ⅲ），直到所有的信息都被使用一次并且仅仅一次。最终结果是每列中条件概率行的数目。

Ⅴ在每列中，将前概率和所有条件概率相乘，得出联合概率P(Hi). P(E | Hi)，联合概率在列间不一定等于1。

Ⅵ如果仅有两列，联合概率可直接用比值比。联合概率可以被相应调整，得出总和为1的终概率。这可以通过将每个联合概率除以所有联合概率的总和Σ[P(Hi). P(E | Hi)]来获得。

使用连锁分析程序来计算遗传风险

乍一看，设计为计算优势对数评分的程序也可用于计算遗传风险似乎令人惊奇，但事实上，这两者是有着密切联系的（图18.16）。在一定的数据和假设的情况下，连锁分析程序是计算系谱可能性的通用工具。为计算连锁可能性，我们计算如下比值：

$$\frac{\text{数据的似然性} \mid \text{连锁，重组值}\ \theta}{\text{数据的似然性} \mid \text{无连锁}\ (\theta=0.5)}$$

为估计先症者携带疾病基因的风险，我们计算如下比值：

$$\frac{\text{数据似然性} \mid \text{先证者是携带者，重组值}\ \theta}{\text{数据似然性} \mid \text{先证者不是携带者，重组值}\ \theta}$$

如框18.4一样，垂直线 | 表示：‘在假设……条件下’

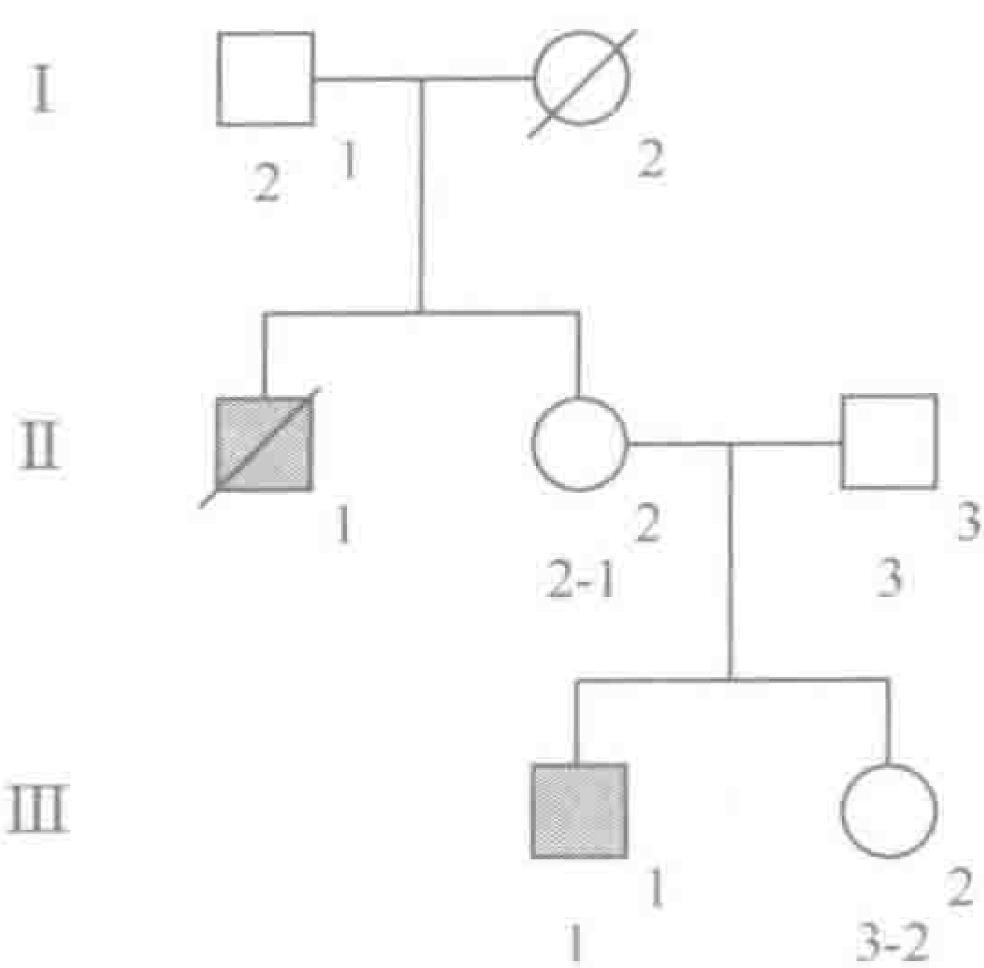


图 18.15　遗传风险的 Bayes 计算

假设：Ⅲ-2 是	携带者	非携带者
前概率	1/2	1/2
条件概率（1）：DNA 结果	0.05	0.95
条件概率（2）：肌酸激酶检测数据	0.7	1
联合概率	0.0175	0.475
最终概率	0.0175/0.4925	0.475/0.4925
	=0.036	=0.964

Ⅲ-2 想知道她是 DMD 携带者的风险，她的兄弟Ⅲ-1 和舅父Ⅱ-1 患有 DMD。血清肌酸激酶检测（DMD 携带者通常具有的亚临床肌肉损伤的标志物）显示携带者与非携带者的可能性是 0.7∶1。与 DMD 基因有平均 5%重组的 DNA 标记给出的基因型分析如图。根据框 18.4 的原则，风险计算给出她是携带者的总体风险是 3.6%。

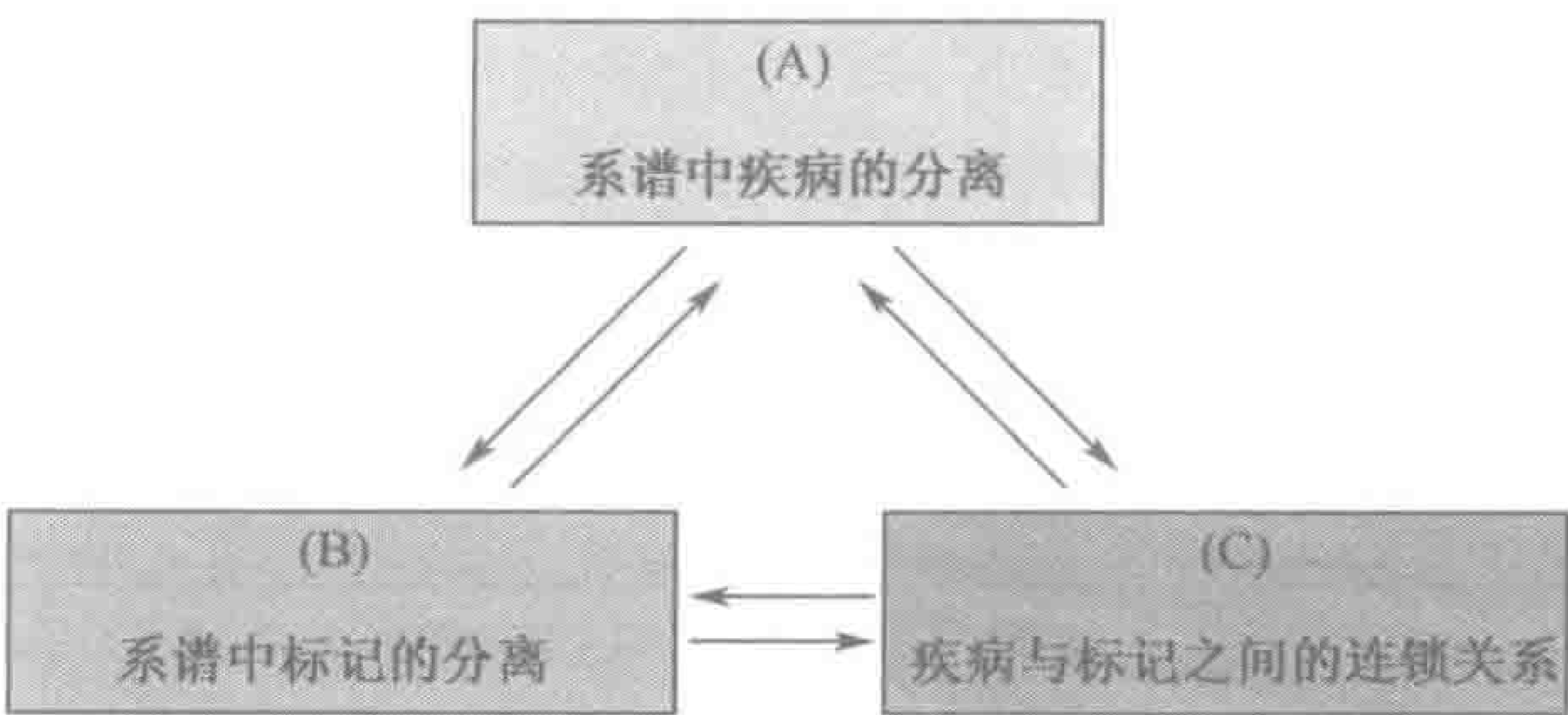


图 18.16　连锁分析程序用于计算遗传风险

已知其中任何两者的信息，可以计算第三者的信息。对于连锁分析，已知（A）和（B），计算（C）。对于计算遗传风险，已知（B）和（C），计算（A）。

18.5.4　杜兴肌营养不良的特殊问题

在杜兴肌营养不良实验室诊断中，会遇到多种问题。所幸的是，该病的突变类型中，有 2/3 为缺失，这在男性中容易识别（图 18.7），虽然在女性中缺失识别具有挑战性。重复突变在男性和女性中都难以发现，并且毫无疑问诊断率较低。点突变中30%～50%构成诊断的主要难题。对于如此大的一个基因（2.4Mb，79 个外显子），进行点突

	受累者	未受累者
筛查试验阳性	a	b
筛查试验阴性	c	d

筛查试验灵敏度 =a/(a + c)
筛查试验特异度 =d/(b + d)
阳性预测值 =a/(a + b)

图 18.17 筛查试验的灵敏度和特异度

变扫描是十分困难的，因此，通常使用基因示踪技术。然而，因为 DMD 基因存在一个非常高的重组率，因此在基因示踪时会出现特殊的问题。甚至基因内标记与疾病也有平均 5％的重组率，因此，如图 18.14 所示，使用旁侧标记应审慎。

但问题并未就此结束，还存在一个高发的新突变。在框 4.7 中突变-选择平衡计算表明：对于任何一种致死性 X-连锁隐性疾病（f＝0）的情况中，1/3 的病例为新突变。因此，一个散发的 DMD 男孩的母亲只有 2/3 的几率是携带者。这会有两种不好的结果：

- ▶ 这使得发病风险的计算更加复杂，而这种计算对解释基因示踪结果是必需的。有兴趣的读者可参考 Bridge（进一步阅读）的书，查看示例计算。
- ▶ 如图 4.8 所示，DMD 系谱中第一个突变携带者通常是嵌合体（男性或女性）。这也会引出更多的问题既有风险计算也有解释直接检测结果的问题。

这些因素，以及该疾病特别令人痛苦的临床病程，家系中高再现风险和群体中 DMD 高发率意味着 DMD 对于遗传服务提供者来说可能是所有疾病中最为困难的。

18.6 群体筛查

群体筛查一般应在具备直接检测突变的能力之后实施。从传统上讲，筛查与诊断是有区别的。筛查是确定高风险人群，然后再做确定性的诊断实验。DNA 检测则相当不同，因为没有单独的筛查实验和诊断实验。不论采用什么技术，开展群体筛查检测的方案仍需要满足相同的标准（表 18.7）。

表 18.7 群体筛查计划的要求

要求	举例及注释
阳性结果必须有所用	▶ 预防性治疗，如 PKU 的特殊饮食
	▶ 囊性纤维化携带者筛查中生殖史回顾和选择
整个计划必须在社会和伦理方面可接受	▶ 受试者必须获得知情同意
	▶ 没有咨询的筛查是不可接受的
	▶ 不能施加压力终止受累的妊娠
	▶ 筛查不能被认为具有歧视性
检测必须具有高灵敏度和特异度	▶ 许多假阴性的检测降低受试者对计划的信心
	▶ 即使能被进一步的确定性诊断实验排除，有许多假阳性的检测也会在正常人群中引起不可接受的高度焦虑
计划的益处必须超过支出	▶ 以无效的方式使用有限的卫生保健预算是不道德的

18.6.1　可接受的筛查程序必须满足特定的标准

筛查应该达到什么目的呢？

任何筛查计划最重要的一项功能是得到一些有用的结果。除非能使他们对这个风险采取什么措施，突然告诉受试者他们存在某些不良风险是根本不能被接受的。筛查乳腺癌或心脏病易感基因的计划必须严格按照这个标准进行评估。对亨廷顿病预测也许似乎打破了这个原则，但是这只是提供给那些已经知道他们具有亨廷顿病高风险的人群，以及那些苦于不确定是否患病的人们，他们要求预测性的检测，并且坚持这样做，尽管在咨询中所有的不利因素都会被指出。

在理想的状态下，有用的筛查结果是治疗，如筛查新生儿苯丙酮尿症。如果能极大地改善预后，增加的医疗监护就是一个有用的结果。过分热衷于筛查的一个大风险就是它能把健康人变成患者。一个特殊的例子就是对携带者的筛查，其结果就是可能避免受累儿童出生，那些不愿意接受产前诊断和终止受累妊娠的人们不会把它视为筛查的有用的结果，一般情况下不应当筛查，尽管有些夫妻重视简单的知情权。

筛查的伦理框架

由享有盛誉的美国遗传学家、医生、律师和神学家组成的一个委员会讨论了群体遗传筛查的伦理问题。读者可参阅他们非常详细的调查报告（Andrews *et al.*，1994）。这是伦理问题的本质问题，他们没有解决办法，但出台了一些原则。

- 任何计划必须是自愿的，受试者主动决定选择它。
- 计划必须尊重受试者的自主权和隐私权。
- 不要迫使得到阳性结果的人接受任何特殊的做法。例如，在以保险为基础的健康保健的国家中，保险公司施加压力迫使携带者夫妻接受产前诊断和终止受累妊娠，这是不可接受的。
- 信息必须保密。这似乎显而易见，但它会是个难题。我们喜欢设想重型卡车或大型喷气式飞机驾驶员都已接受测试过所有可能的危险。以保险为基础的健康保健体系社会对遗传资料保密性有特殊的问题，因为保险公司辩解说，不增加高风险人群的保险费就是在惩罚低风险人群。

18.6.2　特异性和灵敏度衡量筛查试验的技术特性

与伦理问题相比，群体筛查的技术问题相当简单。筛查检测的效果可以用其灵敏度、特异性和阳性预测值来衡量（图 18.17）。

检测的预计值

也许出乎意料，假阳性检测结果比假阴性结果会造成更严重的问题。即使能被随后的诊断性实验排除，许多人仍然不必要地为假阳性结果担忧，表 18.8 表明除非检测非常普遍疾病的时候，如果一个筛查检测确实有显著的假阳性率，那么预测值就非常低。从这个意义上说，DNA 检测可能特别适合群体筛查，因为与采用任意阈值的生化检测

相比，DNA 检测产生的假阳性非常低。

表 18.8　在实验室中很好的检测在群体筛查中可能无用武之地

患病率情况	被筛查群体中真正阳性数	筛查检测出的真正阳性数	被筛查群体中真正阴性数	筛查检测出的假阳性数	筛查预测值
1/1000	1000	990	999 000	9990	0.09
1/10000	100	99	999 900	9999	0.0098
1/100000	10	10	999 990	10 000	0.001

在实验室试验中，对于 100 个受累和 100 个对照人群，预计有 99%准确性。检测出的阳性数占真正阳性中的 99%，检测出的阴性数占真正阴性中的 99%，该表显示筛查一百万人的结果，所有检测阳性人中绝大多数是假阳性。这样的检测不大可能在社会上被接受，对于任何孟德尔式疾病，经济上也不适合（典型情况下，在 1000 人中这类疾病受累者少于 1 人）。

检测的灵敏度

一项检测必须找出预期目标中合理的一部分（即灵敏度必须高）。虽然，DNA 检测的预测值看起来令人欢欣鼓舞，其灵敏度通常依赖于等位基因杂合性。除非一个疾病异常纯合（表 18.4），否则就不适合检测每个可以预见的突变，特别是在大范围群体筛查计划中。通常仅一组突变被筛查。图 18.18 说明了突变的选择如何影响囊性纤维化携带者筛查计划的效果。

仅检测最常见的突变 F508del 不会是一个可以接受的计划，这是清楚的。更多受累儿童出生于在筛查计划中呈阴性的父母，而不是呈阳性的父母。不论这样的计划在经济上是否值得，从社会角度讲，它也是当然不可以被接受的。很难定义什么是一个可被接受的计划。一个建议集中在“+/－”夫妇（一人为已知携带者，其配偶在所有检测中为阴性），该配偶仍然可能是罕见突变的携带者。Ten Kate 建议可被接受的筛查计划是在筛查前这样“+/－”夫妇的风险不高于一般群体的风险。对于北欧囊性纤维化筛查，这需要灵敏度大约为 95%。

18.6.3　遗传筛查计划的组织

假设所提出的计划从伦理上是可接受的，并且很经济，谁应当被筛查？以下三个例子着重说明一些选择。

新生儿筛查：筛查苯丙酮尿症（PKU）

在英国，所有的婴儿在出生几天后即被筛查 PKU。在家庭随访期间，从婴儿脚后跟收集的血样点在卡片（Guthrie 卡片）上，然后送至中心实验室。通过色谱仪或细菌生长实验检测血中的苯丙氨酸水平。这是一项筛查试验。血中苯丙氨酸水平高于阈值的婴儿还要被召回以作确定性的诊断试验。只有一小部分婴儿最终被证明患苯丙酮尿症。婴儿接受饮食治疗所获的益处弥补了缺乏知情同意的不足（Smith，1993）。

产前筛查：筛查β地中海贫血

在婚前或在出生前的诊所中，通过常规的血液学检查可以检测β地中海贫血的携带者。通过 DNA 分析可以对均为携带者的夫妻进行产前诊断。在英国两类种族具有β地中海贫血高发率：塞浦路斯人和巴基斯坦人。在英国塞浦路斯人很快地接受了这种筛查，而巴基斯坦人则接受得慢。这个比较说明了围绕遗传筛查的复杂社会问题和相关文化背景（Gill and Modell，1998）。尤为重要的是，对塞浦路斯社区长期的研究表明怎样衡量筛查是否成功：不是通过受累胎儿流产的数量，而是通过正常家庭的夫妻数量来衡量。在没有实施筛查之前，许多塞浦路斯携带者夫妻选择不要孩子；现在他们正在利用筛查，并且有正常的家庭（Modell *et al*.，1984）。也许有一天，植入前诊断或胚胎干细胞移植会变成终止不良妊娠的一种替代方法，至少在富裕的国家中富裕的人群中可行。

对携带者的群体筛查：对囊性纤维化的建议

现在在技术上是可行的，经济上也是值得去筛查北部欧洲人群以检测囊性纤维化携带者。在英国的普查表明，大多数的携带者-携带者夫妻将选择产前诊断，珍惜这样的机会，确保他们不会有受累子女。如果治疗变得更加有效，例如采用基因治疗，这个观点也许会改变。

如果一个筛查计划将要开展，必须考虑两类问题。实验室应当检测多少突变，谁应该被检测？由等位基因异质性引出的问题在上面已被讨论（图 18.18）。对于筛查谁的问题，表 18.9 说明了在英国被考虑的一些可能性。正常情况下，在每个国家健康保健组织实施的方式将决定可能性的范围。设对照组的实验性研究初步结果表明：以上方法都没有时常预计的负面影响（增加的焦虑）。

表 18.9　组织囊性纤维化携带者群体筛查的可能方式

筛查的群体	优点	缺点
新生儿	▶ 容易组织	▶ 20 年内不会看到筛查结果
		▶ 许多家庭会忘记筛查结果
		▶ 检测儿童是不道德的
学校学生	▶ 容易组织	▶ 难以合乎伦理地开展筛查
	▶ 在他们开始处朋友前通知他们	▶ 有给携带者带来辱名的风险
内科医生名单上的配偶	▶ 配偶风险一致	▶ 很难控制咨询的质量
	▶ 强调内科医生在预防医学中的作用	
	▶ 允许有作出决定的时间	
产前诊所的妇女	▶ 容易组织	▶ 对携带者有出人意料的效果
	▶ 结果迅速	▶ 配偶可能找不到
		▶ 对实验室时间要求紧
成人志愿者（"造访 CF 中心"）	▶ 很少有伦理问题	▶ 咨询情况不好
		▶ 不能对筛查适用者目标定位
		▶ 资源使用效率低下

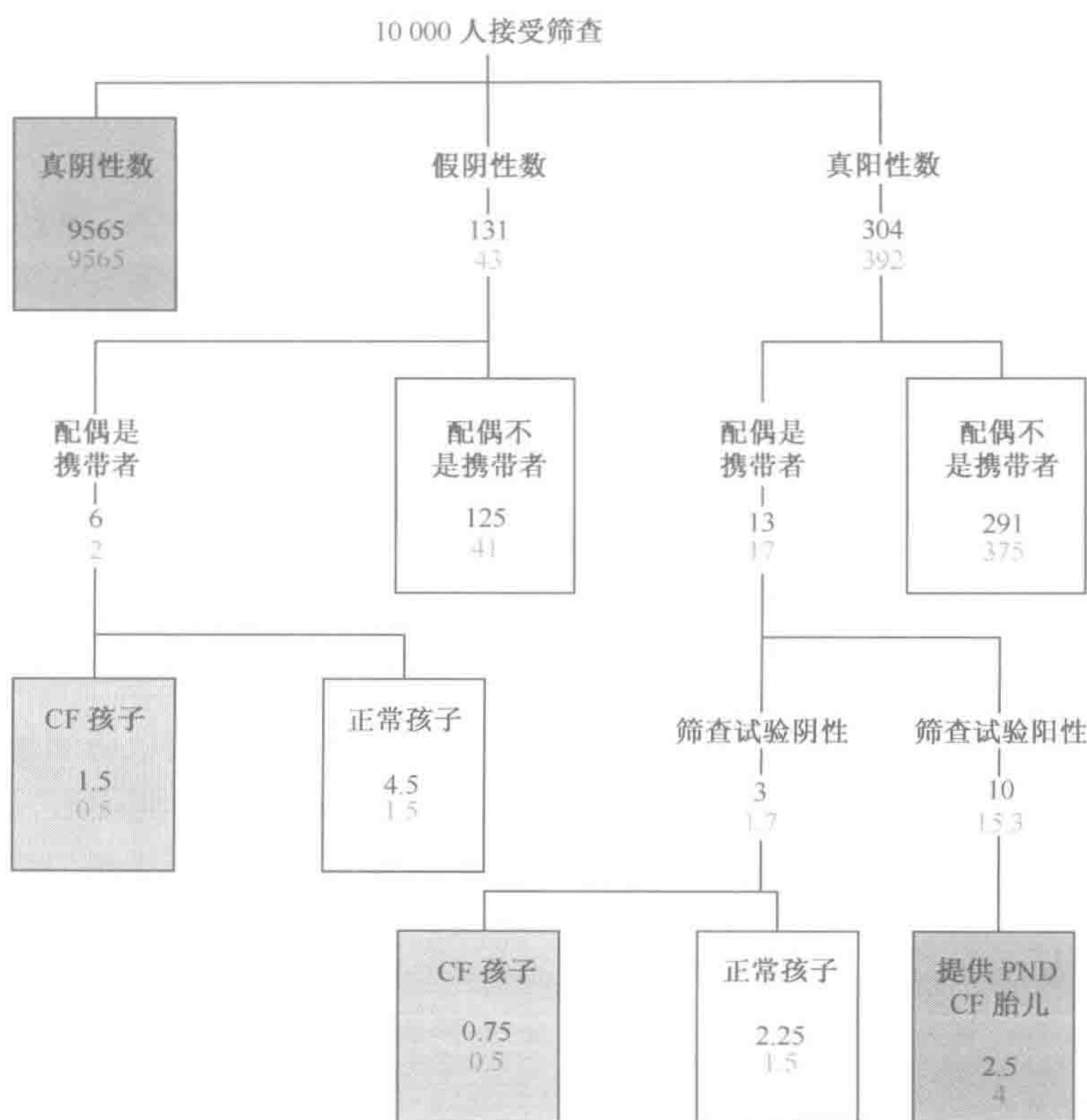


图 18.18　囊性纤维化群体筛查的流程图

筛查 10 000 人的结果 其中 1/23 人为携带者。如果一个人检测为阳性，然后他/她的配偶被检测。黑色数据代表检测 70%囊性纤维化突变（如仅检测 F508del）的筛查结果。浅灰色数据代表检测灵敏度为 90%的筛查结果。深灰色框中数据代表被视为筛查计划成功的例数（不管他们然后采取什么行动），浅灰色框代表失败。PND，产前诊断。

18.7　DNA 图谱可用于识别个体和确定亲属关系

我们采用 DNA 谱这一名词是指一般用于确定身份或亲属关系的 DNA 检测。DNA 指纹专门用于由 Jeffreys 等人（1985）发明的采用多基因座探针的技术。关于整个领域更多的细节，读者应参阅 Evett 和 Weir 所著的书（1998；进一步阅读）。

18.7.1　多种不同的 DNA 多态性已用于 DNA 图谱

采用微卫星探针的 DNA 指纹

这些探针含有由 Jeffreys 等（1985）在肌红蛋白基因中发现的高度变异分散重复序列中的常见核心序列 GGGCAGGAXG，该序列以许多微卫星的形式遍布于整个基因组，基因组中串联重复数目在个体间有差别。当与 Southern 印迹杂交时，探针会显示出个

体特异性的指纹条带（图 18.19）。DNA 指纹给法医学实践带来了一场革命，但现在由于两个问题，它是过时的：

- Southern 印迹操作需要几 μg DNA，可能一百万个细胞中的 DNA 含量与之相等。
- 不可能区分指纹中哪一对条带代表等位基因。因此，当比较两份 DNA 指纹时，调查者通过位置和强度分别匹配每一个条带，不得不把凝胶条带上连续可变的距离分成许多区块。在同一区块内的条带注定要匹配，那么，比如 10/10 条带相匹配，则嫌疑犯而不是人群中随机一个人是样品来源的可能性是 1：P^{10}，在这里，P 是指随机一个人的条带与指定一条带相匹配的几率（我们已进行了简化：假定对于每个区块，P 是一样的并且忽略对条带位置和强度匹配的要求）。即使 P=0.2，P^{10} 仅仅为 10^{-7}。以同样的划分区块的标准判定两个指纹的匹配和计算 P 值，这是绝对必要的。在一定范围内，标准可能具有任意性，但它们必须保持一致。

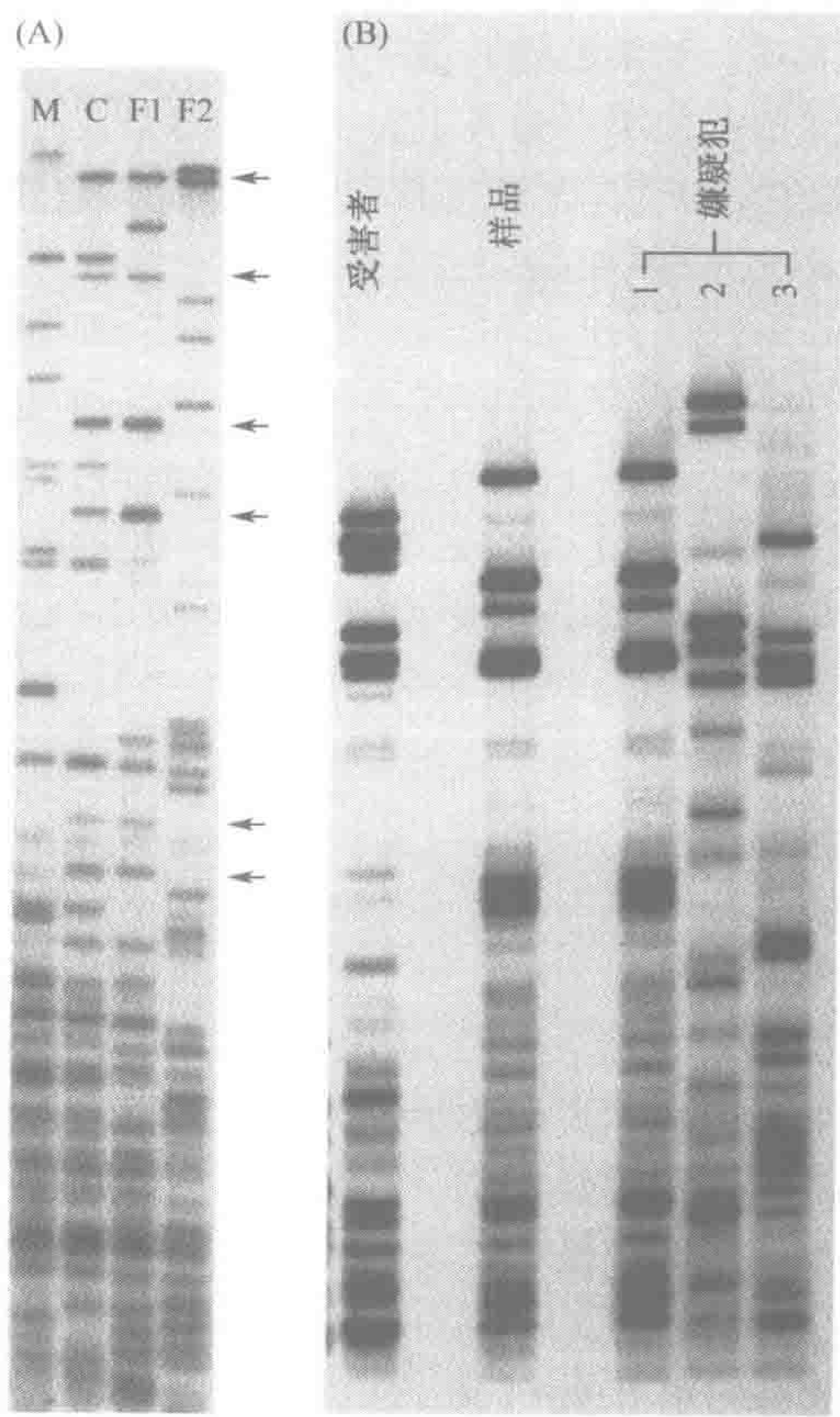


图 18.19　DNA 指纹在法律和法医学上的应用

虽然现在已经被采用多种单一基因座多态性的图谱所代替，DNA 指纹曾给法医学实践带来了一场革命。(A) 父权检测 显示的指纹来自母亲（M），孩子（C）及两个可能的父亲（F1，F2）。F1 DNA 指纹含有在孩子中找到的所有父源条带，而 F2 的 DNA 指纹仅含有一条父源条带。(B) 一个强奸案例。嫌疑犯 1 的 DNA 指纹与受害者阴道拭签上精液样品 S 完全匹配，以此为证据，嫌疑犯 1 被控强奸和有罪。图片由英国 Oxfordshire，Abingdon 细胞标记诊断室惠赠。

用微卫星标记的 DNA 谱

单一基因座 DNA 谱采用微卫星多态性（节 9.4.3），通常采用可由 PCR 分型的三或四核苷酸重复。在理论上，即使犯罪现场遗落的一个细胞也可以进行分型。通过精确的重复数目，等位基因可以明确地予以确定，避免了划分区块的问题。如果群体中每个等位基因基因频率已知，父权概率或嫌犯不是强奸者的概率可被精确计算得出。10～15 个非连锁高度多态基因座组成的基因型在某种程度上具有独一无二的确定性。有些微卫星的重复单位内的微小差异可以用于区分几乎无限多样性的等位基因，所以单一基因座基因型可能足以识别个体（Jeffreys *et al*.，1991），但这种 MVA 分型方法仍然是一种研究工具。

Y 染色体和线粒体多态性的用途

为追溯死者的亲属关系，Y 染色体和线粒体 DNA 多态性特别有用，因为在这两种情况下，个体从一个确定的祖先那里遗传了完整的基因型。有趣的一个例子是通过将挖掘的遗体 DNA 谱与活着的远亲 DNA 谱进行比较，鉴定出遗骨是在 1917 年被布尔维什克杀害的俄罗斯沙皇及其家人（Gill *et al*.，1994）。

18.7.2　DNA 图谱可用于确定双生子的合子型

在研究非孟德尔式性状（章 15）及有时在遗传咨询中，知道一对双生子是单卵双生（MZ，完全相同）还是双卵双生（DZ，兄弟关系）是重要的。传统方法依赖于评价表型相似形或出生时绒毛膜的情况（包含在同一绒毛膜中的双生子总是单卵双生，虽然，反过来这是不正确的）。因为非常相似的双卵双生子被错误地统计为单卵双生，而非常不同的单卵双生子被错误地归类为双卵双生，因此，合子型确定的错误会系统地放大对非孟德尔性状遗传率的估计。

遗传标记提供了一个更为可靠的合子型检测方法。Race 和 Sanger 总结了众多的、但现在已过时的采用血型区别双生子的文献（进一步阅读）。DNA 谱如今是一个可供选择的方法。Jeffreys 指纹探针提供了一种直接的印象：来自单卵双生子的两份样品看起来是一份样品加了两次，而来自双卵双生子的两份样品则表现出差别。当采用单一基因座标记时，如果双生子具有同样的基因型，那么就应该计算每一个基因座上，双卵双生子基因型相似的概率。如果双亲已被确定基因型，这就要遵循孟德尔法则；否则，双卵双生子基因型相同的概率，必须以每一个可能的双亲婚配并以群体基因频率计算所得的同样婚配的概率加以权重来计算。每一个（非连锁）基因座上的概率相乘，得出采用所有标记时，双卵双生子是同一基因型的总体概率 P_1。双生子是单卵双生的概率是：

$$P_m = m/[m + (1 - m)\ P_1]$$

这里 m 代表单卵双生子在群体中的比例（对于同性别双生子，该值为 0.4 左右）。在 Vogel 和 Motulsky 的附录 4 中有示例计算（进一步阅读）。

18.7.3　DNA 图谱可用于反驳或建立父权

排出父权相当简单，如果孩子有一个标记等位基因不存在于其母亲或者假定的父亲

中，那么除了新突变外，该假定的父亲不是孩子生物学上的父亲。确定父权在理论上是不可能的，永远不能证明世界上不存在另外一个人，他能够给予这个孩子这组特殊的标记等位基因。人们所能做的是确立非父权关系的概率，该概率低到足以满足法庭的需要，或如果可能，低到满足假定父亲的需要。

DNA 指纹探针广泛地用于此目的（图 18.19）。正如以上解释的那样，条带必须根据一种任意的但却一致的原则进行区块分类，以确定是否孩子的每一条非母源条带与假定父亲的条带相符。单一基因座微卫星使人们能更为明确地计算可能性（图 18.20）。如果所有条带都吻合，一组十个非连锁高度多态的单一基因座标记能以绝对优势支持父权。

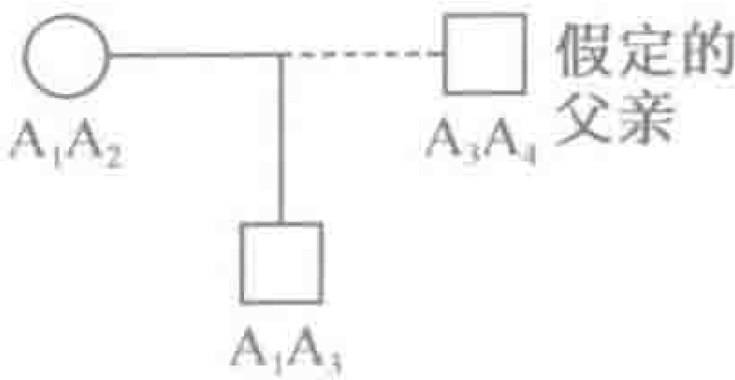


图 18.20　用单一基因座标记进行父权检测

假定的父亲是真实父亲而不是群体中随机成员的可能性是 1/2：q_3，此处 q_3 是 A3 的基因频率。如果父权身份未被排除，将采用一系列 n 个非连锁标记，可能性将是 $(1/2)^n$：q_A. q_B. q_C... q_N。

18.7.4　DNA 图谱是法医调查的强大工具

法医目的的 DNA 谱遵循与父权鉴定相同的原则。对犯罪现场材料（血污，毛发或强奸受害者阴道拭签）进行分型并与嫌犯 DNA 样品比对。DNA 谱最强有力的应用之一是通过证明嫌疑犯不是犯罪者，预防审判不公。无论有任何相反的详细证据，如果样品不匹配，嫌疑犯将被排除。

如果基因型确实全部匹配，法庭需要知道罪犯即是嫌疑犯而不是群体中随机一个成员的可能性。当然，如果罪犯是嫌犯的兄弟或他的同卵孪生兄弟，可能性会相当不同。法庭上 DNA 证据的命运提供给我们深刻洞察科学和法律文化差别的机会。假设嫌疑犯 DNA 谱与犯罪现场样品相匹配，理性地采纳 DNA 资料仍然面临至少三个障碍：

▶陪审团可能只是不相信，或可能选择去忽略 DNA 资料，正如明显发生在辛普森审判案中那样 [见 Weir（1995）对案例的精彩报道]。他们可能认为定罪性的 DNA 是人为设置的。

▶ 不审慎的律师可能试图将陪审团引入错误的可能性辩论中，即所谓起诉者的误区。这包括将假设样品相匹配，嫌疑犯是清白的概率与假设嫌疑犯是清白的，样品匹配的概率相混淆。陪审团应当考虑第一种概率，而不是第二种概率。如框 18.5 所示，两者差别很大。

▶ 对一些以 DNA 为基础的概率计算原则，可能会出现反对意见：

（Ⅰ）倍增原理，即总体概率可通过每个等位基因或基因座的概率相乘得到，它依据基因型是独立的这一假设。如果群体实际上是由在生育上隔离的群体组成，每个人基因型在该群体中相当稳定而在群体间差别很大，这种情况下概率的计算就会出现误导。这是非常严重的，因为依据倍增原理可得出异常确定的可能性。

（Ⅱ）对于单一基因座标记，概率取决于基因频率。DNA 谱实验室保存有基因频率数据库，但基因型频率是否是在与待检案例相关的种族群体中确定的？

从极端上讲，基因型独立性的争论提示 DNA 证据可能识别出罪犯属于一个特定的种族，但不能表明该种族中哪一个人犯这项罪。这些问题已受到广泛的争议，特别是在

美国法庭。论点原则上是正当的，问题是它是否在实践中发挥足够的作用，已经很清楚一般情况下不会。看到持反对意见专家给出正确判定与错误判定差别在百万倍（10^5 ：1对 10^{11} ：1），如果法庭仍裁定 DNA 证据是毫无希望地不可信，而是依靠目击证据（正确识别的可能性 50：50），这是具有讽刺意味的。

框 18.5　起诉者的误区

嫌疑犯的 DNA 谱与犯罪现场的样品吻合，这能给他定罪吗？考虑以下两种不同的概率：

▶ 假设样品匹配，嫌疑犯是无罪的概率；

▶ 假设嫌疑犯是无罪的，样品匹配概率。

采用 Bayesian 标记法（框 18.4），M=匹配，G=嫌疑犯有罪，I=嫌疑犯无罪，第一个概率是 $P_I \mid M$，第二个概率是 $P_m \mid I$。原告的谬误包括辩论相关的概率是 $P_m \mid I$，但事实上它是 $P_I \mid M$。以下的计算表明两种概率是如何的不同。

如果嫌疑犯是有罪的，样品必须匹配：$P_m \mid G=1$。让我们设想，群体遗传学观点表明：随机选择的任何一个人具有与犯罪现场样品相同 DNA 谱的可能性是 $1/10^6$，即 $P_m \mid I=10^{-6}$，设想罪犯可能是群体中 10^7 个男性中的一个人，如果没有其他证据表明他是罪犯，那他仅仅是群体中随机的一员，他有罪的前概率（在考虑 DNA 证据前）是 $P_G=10^{-7}$. 他是无罪的前概率是 $P_I=1-10^{-7}\approx 1$，Bayes 原理告诉我们：

$$P_I \mid M=(P_I.\ P_m \mid I)/[(P_I.\ P_m \mid I)+P_G.\ P_m \mid G]$$
$$=10^{-6}/(10^{-6}+10^{-7})$$
$$=1.0/1.1$$
$$=0.9$$

我们已经看到

$P_m \mid I=10^{-6}$，差别相当大！

法庭通过忽视极具说服力的 DNA 证据来欺骗自己，但是这个计算也表明当没有其他不利于罪犯的证据时，DNA 证据本身不能够可靠地给某人定罪，除非 $P_m \mid I$ 远远低于 10^{-6}，在一个大型城市实施筛查所有男性以找出强奸者的计划时，应当考虑这一点。

（周助人　译）

进一步阅读

Bridge PJ (1997) *The Calculation of Genetic Risks – Worked Examples in DNA Diagnostics*, 2nd Edn. Johns Hopkins University Press, Baltimore, MD.

Clinical Molecular Genetics Society Best Practice Guidelines for Molecular Genetics Services. http://www.cmgs.org

Cotton RGH, Edkins E, Forrest S (eds) (1998) *Mutation Detection: a Practical Approach*. Oxford: IRL Press.

Elles RG, Mountford R (eds) (2003) *Molecular Diagnosis of Genetic Diseases*, 2nd Edn. Humana Press, Totowa, NJ.

Evett IW, Weir BS (1998) *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer.

Race RR, Sanger R (1975) *Blood Groups in Man*, 6th Edn. Blackwell, Oxford.

Vogel F, Motulsky AG (1996) *Human Genetics*, 3rd Edn. Springer, Berlin.

参考文献

Andrews LB, Fullarton JE, Holtzman NA, Motulsky AG (1994) *Assessing Genetic Risks – Implications for Health and Social Policy*. National Academy Press, Washington, DC.

Armour JAL, Sismani C, Patsalis PC, Cross G (2000) Measurement of locus copy number by hybridisation with amplifiable probes. *Nucl. Acids Res.* **28**, 605–609.

Collins JS, Schwartz CE (2002) Detecting polymorphisms and mutations in candidate genes. *Am. J. Hum. Genet.* **71**, 1251–1252.

Fakhrai-Rad H, Pourmand N, Ronaghi M (2002) Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. *Hum. Mutat.* **19**, 479–485.

Gill P, Ivanov PL, Kimpton C *et al.* (1994) Identification of the remains of the Romanov family by DNA analysis. *Nature Genet.* **6**, 130–135.

Gill PS, Modell B (1998) Thalassaemia in Britain: a tale of two communities. Births are rising among British Asians but falling in Cypriots. *Br. Med. J.* **317**, 761–762.

Jeffreys AJ, Wilson V, Thein LS (1985) Individual-specific fingerprints of human DNA. *Nature* **314**, 67–73.

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. *Nature* **354**, 204–209.

Keen J, Lester D, Inglehearn C, Curtis A, Bhattacharya S (1991) Rapid detection of single base mismatches as heteroduplexes on Hydrolink gels. *Trends Genet.* **7**, 5.

Modell B, Petrou M, Ward RH *et al.* (1984) Effect of fetal diagnostic testing on birth-rate of thalassaemia major in Britain. *Lancet* **ii**, 1383–1386.

Monforte JA, Becker CH (1997) High-throughput DNA analysis by time-of-flight mass spectrometry. *Nature Med.* **3**, 360–362.

Nickerson DA, Kaiser R, Lappin S, Stewart J, Hood L (1990) Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl Acad. Sci. USA* **87**, 8923–8927.

Sarkar G, Yoon HS, Sommer SS (1992) Dideoxy fingerprinting (ddF): a rapid and efficient screen for the presence of mutations. *Genomics* **13**, 441–443.

Sheffield VC, Beck JS, Nichols B, Cousineau A, Lidral A, Stone EM (1992) Detection of multiallele polymorphisms within gene sequences by GC-clamped denaturing gradient gel electrophoresis. *Am. J. Hum. Genet.* **50**, 567–575.

Sheffield VC, Beck JS, Kwitek AE, Sandstrom DW, Stone EM (1993) The sensitivity of single strand conformational polymorphism analysis for the detection of single base substitutions. *Genomics* **16**, 325–332.

Smith I and MRC Working Party on Phenylketonuria (1993) Phenylketonuria due to phenylalanine hydroxylase deficiency: an unfolding story. *Br. Med. J.* **306**, 115–119.

Thomassin H, Oakeley EJ, Grange T (1999) Identification of 5-methylcytosine in complex genomes. *Methods* **19**, 465–475.

Tõnisson N, Zernant J, Kurg A *et al.* (2002) Evaluating the arrayed primer extension resequencing assay of TP53 tumor suppressor gene. *Proc. Natl Acad. Sci. USA* **99**, 5503–5508.

van der Luijt R, Khan PM, Vasen H *et al.* (1994) Rapid detection of translation-terminating mutations at the adenomatous polyposis coli (*APC*) gene by direct protein truncation test. *Genomics* **20**, 1–4.

Wallace AJ, Wu CL, Elles RG (1999) Meta-PCR: a novel method for creating chimeric DNA molecules and increasing the productivity of mutation scanning techniques. *Genet. Test.* **3**, 173–183.

Weaver TA (2000) High-throughput SNP discovery and typing for genome-wide genetic analysis. *New Technologies for the Life Sciences: a Trends Guide* (Supplement issued for *Trends* journals) 36–42.

Weir BS (1995) DNA statistics in the Simpson matter. *Nature Genet.* **11**, 365–368.

White S, Kalf M, Liu Q *et al.* (2002) Comprehensive detection of genomic duplications and deletions in the DMD gene, by use of multiplex amplifiable probe hybridization. *Am. J. Hum. Genet.* **71**, 365–374.

第 19 章　后基因组计划：功能基因组学、蛋白质组学和生物信息学

本 章 内 容

19.1　功能基因组学概述

19.1.1　没有功能注释，从人类基因组计划结构阶段中所获取的信息使用受限

人类基因组计划结构阶段的最终目的是提供完整的基因组序列，即其全部三十亿碱基！正如第 8 章讨论的，2001 年 2 月发表了两个序列草图（国际人类基因组测序协作组，2001；Venter *et al.*，2001），而完整序列是在 2003 年中期获得的。尽管这无疑是生物学空前未有的成就，但是序列本身并不是额外提供信息的。本质上，它们仅是长串的四个字母 A、C、G 和 T。为了明确基因组序列是如何帮助构造一个发挥作用的人，必须挖掘这些序列数据以提取有用的信息。

在任何基因组计划中，测序后的首要任务之一就是鉴定所有的基因，因为这些是基因组的主要功能单位。对于人类基因组计划来说，在先前的研究中已经鉴定了许多基因。预测其他的基因则是根据它们的结构，及其在其他基因组中的保守性，或者它们与表达序列标签（EST）匹配的事实（节 8.3.5）。人类基因的确切数目仍然未知，但是现在大多数估计赞同大约 30 000 的数字。因此距离组装一个完整的、精确的人类基因组的**基因目录**（gene catalog），我们还有一段漫长的路。然而，即使能够获得这样一个

资源，它也只不过是一个组成部分的名单。在我们能够开始了解那些组成部分如何构建一个人之前，我们必须了解它们做什么。所以下一个任务就是确定人类基因组中每一个基因的精确功能，这一过程被称为**功能注释**（functional annotation）。

19.1.2　能够在生化、细胞和整个有机体的水平描述个体基因的功能

一个基因的功能实际上是其产物（们）的功能。大多数基因编码蛋白质，但是绝大多数基因（图 9.4）产生非编码的 RNA 分子。可从三个水平划分人类基因产物的功能：

- 生化水平：例如，一个蛋白质可以被描述为一个激酶或一个钙结合蛋白。这很少揭示其在有机体中更广泛的作用。
- 细胞水平：这建立在关于细胞内定位和生物学通路信息的基础上。例如，（我们）可以确定一个蛋白质定位于细胞核且是 DNA 修复所必需的，即使不清楚其精确的生化功能。
- 有机体水平：这可以包括一个基因表达的空间和时间信息及其在疾病中的作用。例如，*HD* 基因产生两个主要的转录物，一个在脑中含量丰富，编码一个叫做亨廷顿的蛋白质。增加此蛋白中谷氨酰胺残基数目并使其超过某一特定限制的突变导致亨廷顿病（节 16.6.4）。在健康人中该基因产物精确的生化和细胞功能仍有待确定。

为了获得一个完整的人类基因功能的图画，所有这三个水平的信息都是必需的。作为一个例子，人类葡萄糖激酶基因的功能性分类显示于框 19.1。功能基因组学一个主要的近期发展是建立了限定词汇表，以描述所有基因组中基因的功能。一个例子是基因本体论系统（基因本体论协作组，2000，2001；http://www.geneontology.org/）。其他有用的系统包括那些基因和基因组的京都百科全书（Kyoto Encyclopedia of Gene and Genome，KEGG；http://www.genome.ad.jp/kegg/）所使用的系统，以及所有系统中最古老的、应用最广泛的系统——用于酶分类的酶学专门委员会系统。

框 19.1　葡萄糖激酶的功能

基因名称：*GCK*

基因组位置：7p15

蛋白质名称：葡萄糖激酶

生化功能：激酶，底物是葡萄糖

细胞功能：葡萄糖代谢，糖酵解途径

有机体功能：在胰腺 β-细胞和肝细胞中特异性表达，葡萄糖调控性胰岛素分泌的主要调节因子，功能丢失性突变导致糖尿病，功能获得性突变可导致高胰岛素血症。

就像许多蛋白质的情况一样，葡萄糖激酶的生化和细胞功能并未揭示其在整个有机体水平的重要调节作用。事实上，人类基因组中编码了其他几个具有相同生化和细胞活性的酶，但每一个酶都有一个独特的表达模式和一个不同的更高水平的功能。反过来，表达模式和疾病数据揭示了葡萄糖激酶调节胰岛素产生和血糖水平的重要性，但并未鉴定其精确的生化活性。

19.1.3　必须在转录物组和蛋白质组水平研究基因间的功能关系

即使基因组中的每一个基因都被鉴定并指派功能，这仍旧不能显示基因产物如何协调制造一个充满生气的人所需的生物活性。一个类似的情况可能是对一部小汽车所有配件详尽的功能描述。这个螺丝可以将发动机固定在底盘上，这是部分操纵系统，这是一个操作指示灯的电动开关，但是实际上汽车是怎样工作的？为了正确评价无数个基因的功能如何联合起来而产生一个人，（我们）必须直接研究基因产物。引人争议的是，在过去的三十年，为了解释生物如何运转，分子生物学的全部目的一直是确定基因功能，并使基因与通路和网络相联系。然而，最近研究基因及其产物的方法清晰地从每次一个的**还原论方法**（reductionist approach）转变为同时研究许多或实际上所有基因产物的**整体论方法**（holistic approach）。这种基因功能的整体分析是**功能基因组学**（functional genomics）的基础。

功能基因组学的中心思想是基因组的表达以产生转录物组和蛋白质组。**转录物组**（transcriptome）是一特定细胞中 mRNA 的完整集合，是转录、RNA 加工和 RNA 更新的复合产品（Velculescu *et al.*，1997）。这是定义**蛋白质组**（proteome）所必不可少的，蛋白质组是一特定细胞中蛋白质的完整集合（Wasinger *et al.*，1995）。重要的是要认识到转录物组和蛋白质组比基因组更复杂。一单个基因能够通过选择性剪接、选择性启动子或多聚腺苷酸化位点的使用，以及像 RNA 编辑一样的特殊加工策略（节 10.3.3），产生许多不同的 mRNA。由这些 mRNA 合成的蛋白质可经各种各样不同的途径进行修饰，例如蛋白水解酶裂解、磷酸化或糖基化。不同于基因组（它在大多数细胞中是相同的），转录物组和蛋白质组是高度可变的。转录、RNA 加工、蛋白质合成以及蛋白质修饰都可调节，结果转录物组和蛋白质组在不同细胞类型以及对细胞所处环境变化的反应显著不同。在转录物组和蛋白质组水平的分析提供了一个处于活动的细胞的快照，显示了在一组特定环境条件下所有 RNA 和蛋白质的充裕。通过了解在不同细胞类型中转录物组和蛋白质组最重要的特征，以及研究它们在健康和疾病中如何改变，就有可能构建基因个体功能的巨幅画卷。

19.1.4　高通量分析方法和生物信息学是实现功能基因组学的技术

功能基因组学已经从新型实验策略的全线发展中获益匪浅，得以在整体水平上研究基因的功能。在一些实例中，现存的技术适用于高通量分析。例如，可用于整体基因表达分析的 DNA 芯片和序列抽样方法已经替代了简单的逐个基因杂交技术（节 7.3.2）。在其他的实例中，已经发明了一些全新的技术，如应用酵母双杂交系统（见下文）进行相互作用筛查。因为这些实验产生了大量的数据组，所以生物信息学的支持是必不可少的。新算法的发展和新数据库的建立同实验方法一样，已经推动了功能基因组学的革命。生物信息学是序列和结构的比较、结构和相互作用的造型、整体表达数据分析，以及通过可进入的、用户便利的数据库实现信息共享所必需的。现存的生物信息学技术已修改为新的用途（如系统发生分析所使用的聚类算法，已适用于采集基因表达数据），并且为了特定的应用发展了新技术（如利用质谱数据检索蛋白质数据库的算法）。我们在本章的剩余部分讨论这些新技术及其应用。

19.2　通过序列比较进行功能注释

19.2.1　通过序列比较能够确定暂时的基因功能

应用同源性检索进行功能注释是基因发现的扩展

各种实验方法可用于检测基因组 DNA 中的基因（节 7.2）。然而，由于基因组计划中产生的大量序列数据，最初的注释是利用能够非常快速地加工序列的计算机算法实施的。正如节 8.3.5 讨论的，这种算法要么根据基本原理预测基因的存在，要么通过检索数据库鉴定与已知基因同源的序列。关于这些方法已在框 8.7 中更详细地描述。

同源基因具有相似的序列，因为它们来源于共同的进化祖先。一种进化关系通常预示这两个序列不但在结构上，而且在功能上相关。因此，指定一个新基因功能的最简单的方法是寻找已经被注释的相关序列。一般地，比较在蛋白质水平进行，因为在进化关系上氨基酸序列比核酸序列更保守，所以蛋白质序列比较更灵敏。

基于同源性检索的功能注释的耐用性依赖于许多因素，包括查询序列和任何数据库匹配序列的相似性程度，数据库中已存在的功能性信息的可信度，以及保守序列和保守结构符合保守功能的程度。一个特定的查询序列可以得到一些具有不同相似性程度的匹配序列。在一些实例中，可能沿着它们的全长比对匹配序列，这表明仅仅点突变的积累就使序列发生差异（图 19.1）。如果序列非常相似，则它们可能代表同源基因，在不同的物种中执行相同的功能，而由于物种形成积累了突变。就像在框 12.3 中谈到的，这样的基因称为**种间同源**（ortholog），一个例子可能是人类和绵羊 β 珠蛋白基因。基于种间同源的功能注释能够非常准确。较低的相似性程度可能提示基因是同源的，但在功能关系上有差异。这样的基因称为**种内同源**（paralog），源自基因组内的基因复制和趋异（框 12.3）。人类肌球蛋白基因和 β 珠蛋白基因是种内同源的例子。在这些实例中，生化水平的功能预测是可信的（这两个蛋白都是氧的携带者），但是它们特有的细胞和有机体水平的功能可能非常不同。普遍的事实是结构相似性越大，暗示功能相似性越大，生化功能比细胞和有机体水平的功能更高度保守。



图 19.1　一既定的蛋白质序列的相似性检索可以鉴定同源蛋白质，其序列沿查询序列（Q）全长与查询序列比对

这些序列由于单独的点突变的独自积累而存在差异，点突变可能是氨基酸置换（由白线代表）或小的插入和缺失。一般地，查询序列（Q）与得到序列的差异越大，功能的保守性可能更小。

在许多实例中，数据库检索并不能得到沿查询序列全长与其匹配的结果。反而，在一些看起来根本不相关的蛋白质中鉴定了部分比对（图 19.2）。这反映了蛋白质分子的

模块化特性，以及不同的蛋白质结构域执行不同功能的事实（节 12.1.3）。匹配基因并不仅仅单纯地由于点突变的积累造成差异，也可以由于更复杂的事件造成差异，例如基因与基因片段间的重组导致外显子混编（节 12.1.3）。涉及血液凝固的人类蛋白质为此过程提供了一个有用的例子（Kolkman and Stemmer，2001；图 12.3）。

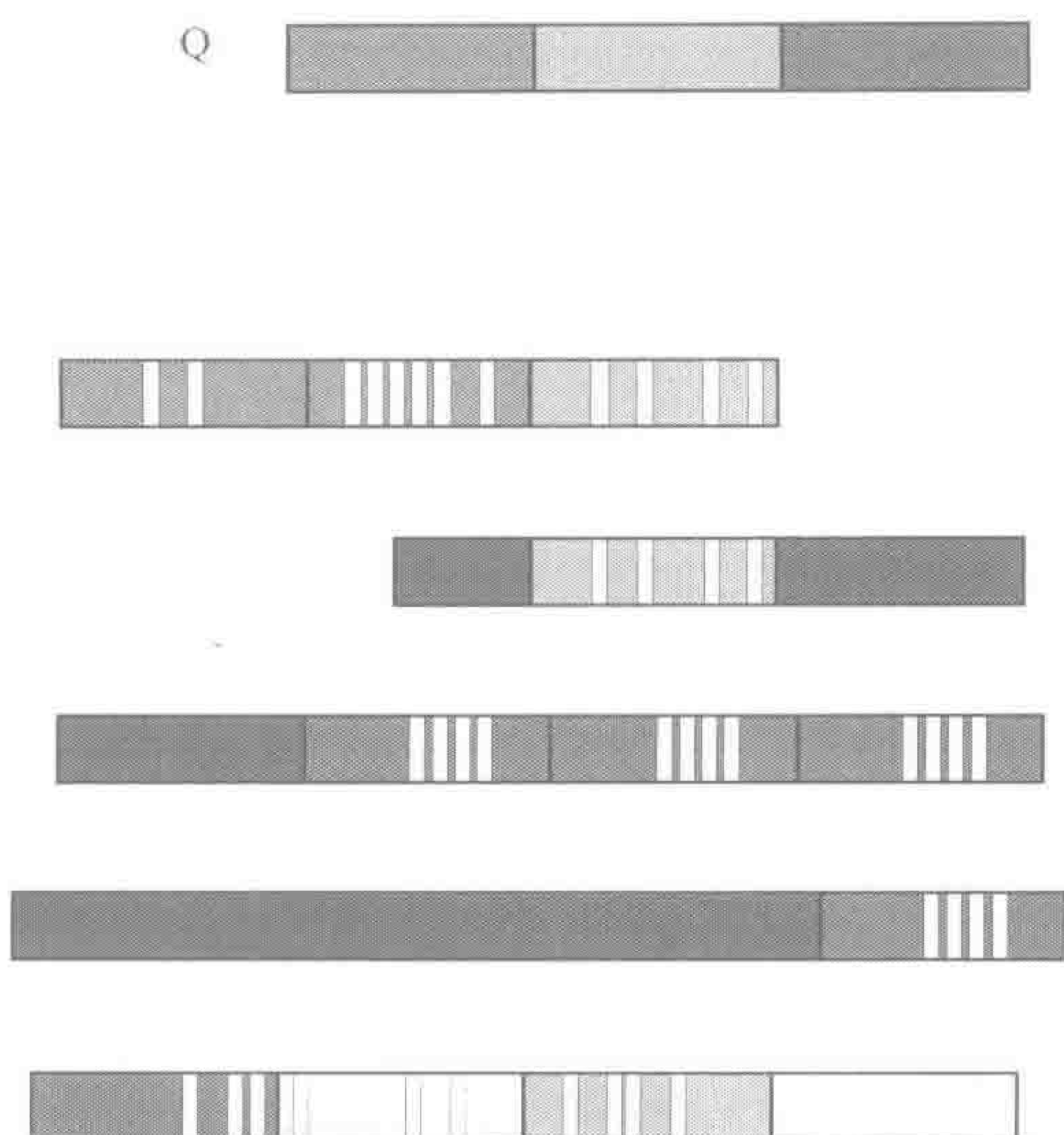


图 19.2　一给定蛋白质的相似性检索也可以鉴定同源蛋白质，表现出与查询序列部分比对
在这个例子中，查询蛋白质（Q）包含了三个截然不同的结构域，以不同的颜色显示。检索到的序列可以共享这些结构域的一个、两个或全部三个，但也可以具有查询序列中不存在的额外的结构域（黑色方框）。可以是结构域复制，如在第一和第三个查询结果中所见到的。最后的查询结果显示了前突变的特殊情况，此处结构域的顺序发生改变。这通常反映了基因复制，以及随后的基因融合和末端侵蚀。与查询序列相关的点突变的积累以白线显示。

同源性基础上功能注释的几个缺点

基于同源性检索的功能注释的原理是保守的结构总是反映保守的功能。然而，有几种这种假定是不安全的情况（Orengo *et al.*，1999）。

- **低复杂性序列**（low complexity sequence）的存在，即存在于许多蛋白质中，具有极度不同功能的序列。例如，跨膜结构域，二聚体化结构域；
- **多功能序列**（multifunctional sequence），即在不同蛋白质中执行不同功能的序列。例如，在六种不同类型酶的催化位点，以及细胞黏附分子 neurotactin 中发现了形成一个 α/β 水解酶折叠的序列。
- **基因复原**（gene recruitment），即现存基因产物一个新功能的获得。例如，涉及普通代谢过程的许多酶已经复原为**晶体蛋白**（crystallin），即眼晶状体中的折射蛋白。这仅仅通过修饰它们的表达水平而实现，而其他的复原机制包括改变蛋白质形成复合物的途径和改变它们的细胞内定位。

同源性检索的另一个缺点是注释必须建立在其他人的实验和阐述基础之上。然而，

所有的数据库，无论多么仔细地管理，仍旧含有一显著比例的错误。将一个新基因的功能建立在这样的数据基础上，不仅不正确，而且还谬种流传（Brenner，1999）。

19.2.2　一致性检索方法能够扩大已鉴定的同源关系的数量

应用了 BLAST 家族算法的标准同源性检索方法（表 8.2 和框 8.7）适用于检测与其全长或者在一个或几个结构域密切相关的蛋白质序列。然而，当序列水平的相似性低于 30%～40%，这些算法就变得不健全，而且许多进化关系都被忽略。

改进同源性检索成绩的一个方法是使用一致性检索方法，如 PSI-BLAST（**位置特异性重复 BLAST**，position-specific iterated BLAST），它应用了建立在序列分析基础上的重复检索（Altschul *et al.*，1997）。原理见图 19.3 所示。这个过程的开始如同正常的 BLAST 检索，但是得到的初级匹配序列整合并形成一个代表性的序表（profile），随后用于第二轮检索。任何额外的匹配序列均与序表整合，这一过程可重复至预先确定的重复次数，或者直到没有鉴定新的匹配序列为止。此方法已显示鉴定了三倍于标准同源性检索的进化关系。通过匹配模式，由远相关蛋白比对产生的序表，以及符合功能结构域的短保守基序或更长序表的衍生物，可获得更高的灵敏度（Eddy，1998）。已经产生的许多二级序列数据库，含有来源于初级序列数据库的结构域基序或序表（表 19.1）。可以使用查询序列检索这些数据库以便鉴定非常远的相关序列中保守的蛋白质结构域。

表 19.1　可用于鉴定保守的元件和蛋白质结构域的蛋白质序列二级数据库。Interpro 是一个有价值的交叉参考系统，允许使用一单个查询序列检索每一个数据库

数据库	内容	URL
PROSITE	与蛋白质家族相关的**序列模式**（sequence pattern）和更长的代表全部完整蛋白质结构域的**序列谱**（sequence profile）	http://ca.expasy.org/prosite
PRINTS，BLOCKS	在多个蛋白质家族中匹配序列高度保守的区域。这些在 PRINTS 中被称为**基序**（motif），在 BLOCKS 中被称为**模块**（block）	http://bioinf.man.ac.uk/dbbrowser/PRINTS http://www.blocks.fhcrc.org
Pfam，SMART，ProDom	蛋白质结构域的集合	http://www.sanger.ac.uk/Software/Pfam http://smart.embl-heidelberg.de/
Interpro	整合来自其他二级数据库的信息的检索工具	http://www.ebi.ac.uk/interpro/

19.2.3　基因组间的相似性和区别指出保守性和功能性的重要序列

从上述的讨论中，很清楚的是来源于不同物种（种间同源）的近相关的基因，能够用于指定不具有特性的基因的精确功能。种间同源通常并不相同，因为在物种形成事件之后，独立的突变在每一进化系中积累。所以，种间同源间的相似性程度提供了一个有用的进化时间的衡量标准，并可用于构建种系进化树（12 章）。**比较基因组学**（comparative genomics）利用基因组间的相似性和区别来得到结构、功能和进化信息（节

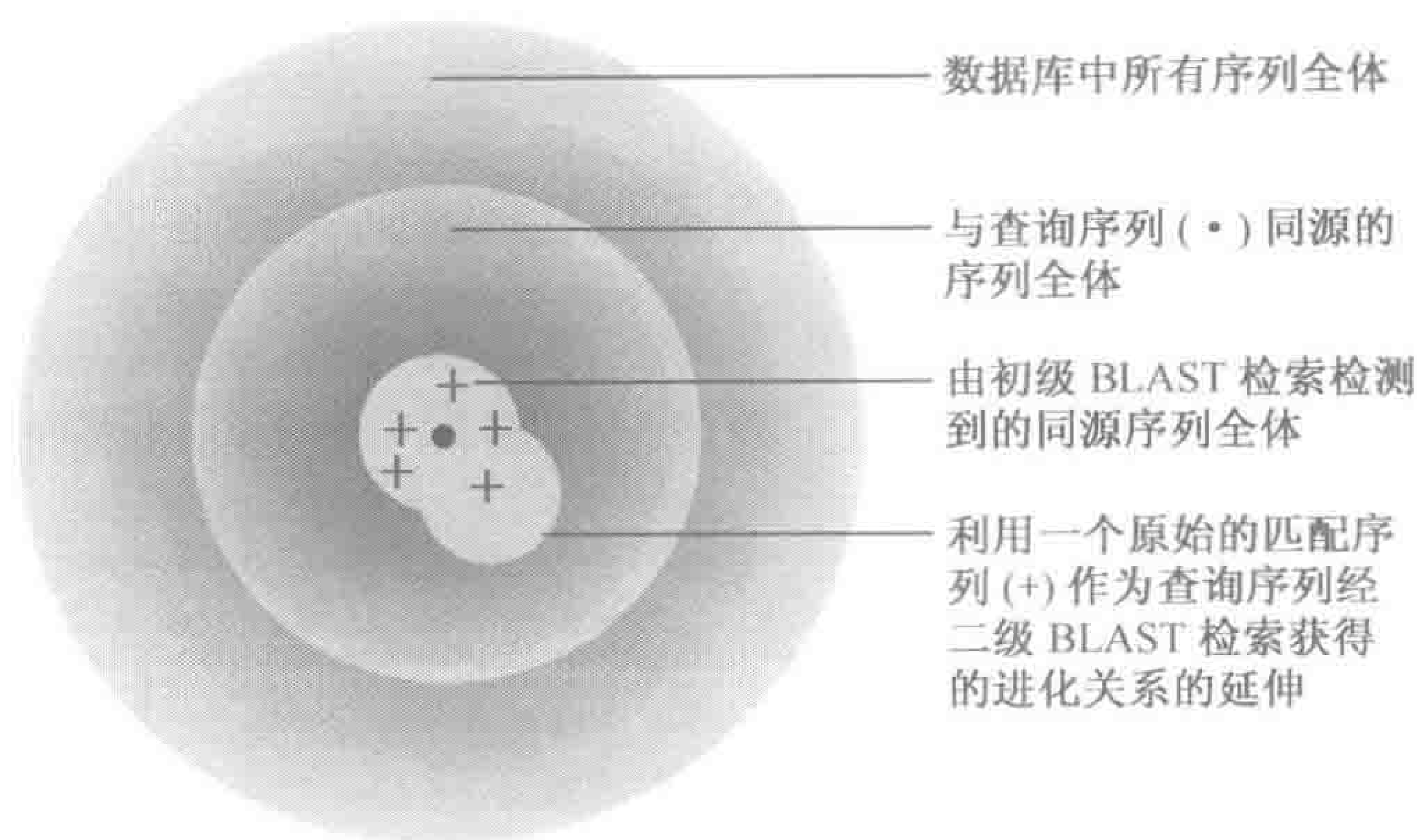


图 19.3　PSI-BLAST 的原理

序列数据库的全部内容由外层的圆圈代表。这些序列的一个子集与查询序列同源（点），尽管进化关系程度与来自中心的序列相距很远。与查询序列相关的一个小的内层圆圈的序列将由标准的 BLAST 检测到。在 PSI-BLAST 中，第一次检索中的每一个匹配序列被编辑成一个序表，它被用于二级检索，以扩大已鉴定的同源序列的数量。这一过程可重复至预先确定的检索次数，或者直到不再发现匹配序列。

12.3.2)。这基于一定的原理，即物种间遗传相似性比基因水平延伸得更深。据说密切相关的物种具有相似的基因和相似的基因组，相似的程度取决于物种的进化差异（Nadeau and Sankoff，1998）。在序列水平，还有在基因组结构的大体水平，相似性显而易见，以至于相关的物种通常显示出**保守同线性**（conserved synteny）（一个保守的基因顺序）。

同线性可适用于基因定位和克隆，因为来自一个物种的图谱信息可以用于定位和克隆另一个物种的基因。至于人类基因组，花费在高密度遗传图构建（第 8 章）的大量尝试，目前在不止一条途径中得到回报，因为人类和其他脊椎动物间保守同线性的证据正用于定位和克隆其他哺乳动物相应的基因。鉴定人类疾病基因的动物种间同源（基因）很有价值，因为这可以使我们了解为什么人类对特定疾病具有易感性，因而有助于这些疾病的预防以及新药开发。例如，大多数哺乳动物对 HIV 不易感。鉴定赋予疾病易感性的人类基因（如 *CCR5*，编码 T 细胞表面的 HIV 共受体）的种间同源体（基因）可能提供新的治疗策略。

比较基因组学的另一个应用是基因调节元件的鉴定。如第 20 章讲到的，正常基因表达所需的启动子和增强子元件的鉴定是一项艰巨的任务，涉及培养细胞中的人工表达实验和转基因动物。一个可能的捷径是比较相关物种的种间同源基因，寻找基因编码区外的最保守的序列基序。只有具有高度保守功能的序列才能在两个基因组中出现，而其他序列将会在进化时间上具有明显的差异（Hardison *et al*.，2000；Werner，2003）。在这个方面，日本河豚鱼红鳍多纪（Takifugu rubripes）提供了一个极好的模型，因为它具有所有脊椎动物中已知的最小基因组（400Mb），但含有大约与人类基因组相同数目的基因（Elgar *et al*.，1996；Aparicio *et al*.，2002）。

19.2.4　比较基因组学能够用于鉴定和描述人类疾病基因的特征

应用同源性检索指定人类基因的功能，通常造成粗略的功能指定，诸如“磷酸化酶”或“跨膜蛋白”。比较基因组学能够用于富集同源性检索提供的信息，因为在相关的基因组中不仅个别蛋白质的功能，而且整个通路、网络和复合物的功能都倾向于保守。

甚至对于相关性远的有机体，这样的相似性可表达于整个有机体水平。例如，据估计大约 30%已知的人类疾病基因在酵母中有同源基因。酵母 *SGS1* 基因编码一个 DNA 解旋酶，与人类基因 *WRN* 同源，该基因的缺陷见于 **Werner 综合征**（Werner syndrome，MIM 2777000）。这是一种早衰性疾病，酶缺陷的个体在他们生命的早期看起来正常，但在生命中期显著衰老，到 30 岁中期形成耄耋老人皱缩的外貌。疾病的常见特征包括动脉粥样硬化、骨质疏松、糖尿病和白内障。疾病的这些症状被认为是在某种程度上与 Werner 细胞在培养基中丧失了像正常细胞一样多次分裂的能力有关，即早期的细胞**衰老**（senescence）。解旋酶活性和细胞衰老间的联系在人类难以检查，但是携带一个没有活性的 *SGS1* 基因的酵母细胞也显示出加速性衰老，为在酵母细胞中开展人类疾病的功能研究提供可能。

人类和动物间基因功能的保守性更大于人类和酵母间基因功能的保守性。超过 60%的人类疾病基因在果蝇和蠕虫中具有相似物，揭示了大约 1500 个基因家族在所有动物中保守的精髓。胰岛素信号通路在人类和线虫中完全保守，所以胰岛素信号受损的蠕虫突变体是**Ⅱ型糖尿病**（type Ⅱ diabete）的有用模型。由于它们微生物样的特性，这些线虫突变体可用几千种可能的药物来筛查，以鉴定使胰岛素不敏感的疾病生理恢复到正常的复合物。秀丽新小杆线虫突变体提供了许多其他疾病模型，包括神经疾病、先天性心脏病和肾脏疾病。我们在下一章更详细地考虑动物疾病模型的应用。

19.2.5　少数顽固基因抵抗由同源性检索进行的功能注释

第一个被完整测序的主要模式生物基因组是酿酒酵母基因组（Dujon，1996）。在获得序列之前，人们普遍认为大多数酵母基因已被实验性地鉴定出来，而基因组序列将最多发现几百个额外的基因。所以当科学家发现酵母基因组有超过 6000 个基因而仅有 30%是先前已知的时候大吃一惊。同源性检索为另外 30%基因提示了功能，尽管注释在它们的应用中变化巨大。在某些情况下，（我们）有可能鉴定生化和细胞功能，但对于许多基因仅能预测一般的生化功能。不管怎么样，这遗留了 40%的基因没有功能指定，这些基因能分为两类（图 19.4）：

- **孤独基因**（orphan gene）：这些是未与数据库中任何其他序列匹配的预测基因；
- **孤独家族**（orphan family）：这些是在数据库中有同源序列，但是同源序列本身功能未知的预测基因。

对于人类基因组，由于基因预测的不确定性使得情况更为复杂。在酵母基因组中，低于 10%的基因预测是不可靠的，但这一数字在人类基因组中更高，可能约为 25%。此外，当许多人类基因已经被鉴定时，预测的结构可能不准确（如，可能有丢失的外显子、错误限定的外显子和虚假外显子，或者邻近基因可能被融合）。考虑到这些情况，

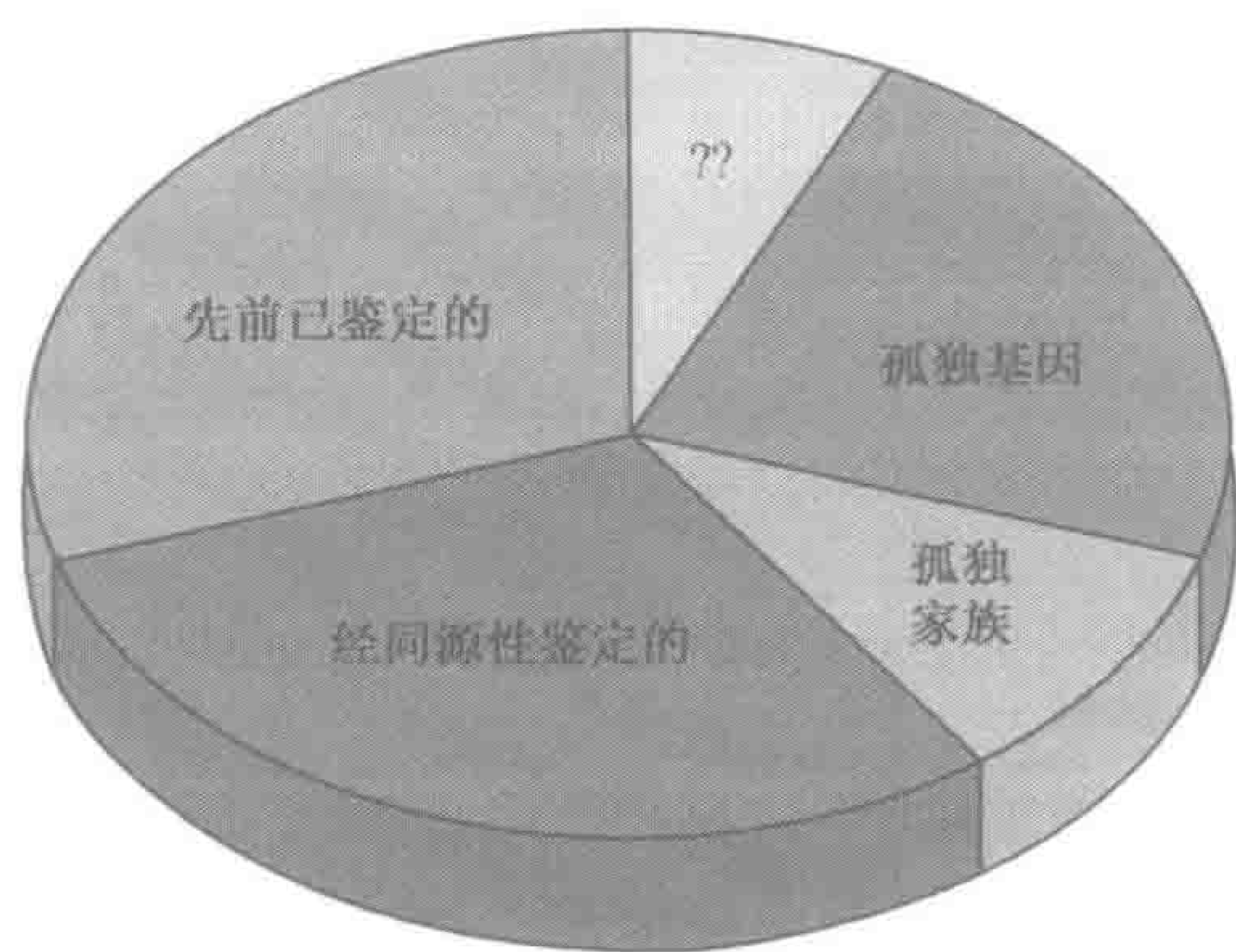


图 19.4 酿酒酵母基因组计划完成后，依据注释状态的酵母基因分布

总体上，6000 个基因中的 30%先前已被鉴定，30%依据同源性检索能指定功能，33%是孤独基因或孤独家族的成员。7%以‘??’代表，是可疑的可读框架。同源性检索基础上的功能注释从充分提供信息的（确定生化和生理功能）、部分提供信息的（确定可能的生化功能）到表面性的（如，可能涉及氮代谢）。

人类基因组计划的结果与从酵母计划中获得的结果惊人地相似。大约 1/3 列出的基因是先前已知的，在 RefSeq 数据库中出现，RefSeq 数据库是一个人类基因转录物高度集中的集合体。另外 1/3 是在与 EST 或其他序列同源的基础上预测的，而剩下 1/3 仅仅基于结构标准而被从头开始预测。在后一类中，一些预测会是假阳性（如与假基因或转座因子匹配），而许多基因会由于从头开始基因预测算法的低敏感性而被遗漏。在先前已鉴定的基因和那些根据同源性预测的基因中，大多数已经至少在生化水平进行了功能注释，而少数属于孤独家族。由从头开始方法预测的基因只有一小部分已确定了功能。大约 60%人类基因经预测含有在二级序列数据库中存在的蛋白质结构域，剩余 40%未确定。在这一类中，可能 10%基因的功能已知，但它们不属于大的蛋白质家族，还有 1/3 的基因根本没有确定功能，是真正的孤独基因。

对于这些孤独基因，同源性检索没有提供功能信息，因此我们必须在其他证据基础上进行功能预测。可用的功能注释方法包括下列方法，将在本章剩余部分讨论：

- ▶ mRNA 表达谱分析
- ▶ 蛋白质表达谱分析
- ▶ 蛋白质结构比较
- ▶ 蛋白质相互作用分析
- ▶ 突变体表型分析或模式生物的表型复制

19.3 总 mRNA 谱（转录物组学）

19.3.1 转录物组分析揭示健康和疾病中基因表达模式的改变如何与细胞的生化活性相协调

表达模式提供了关于个体基因功能的有用的补充信息，也能够凸显它们之间的联系

基于序列和结构比较的功能注释是能够提供信息的，但在许多情况下，它们只提示一个广泛的生化分类，诸如“激酶”或“DNA 结合蛋白”，需要进一步的实验在细胞和整个有机体水平上确定基因功能，研究如何协调个体基因产物活性。一个基因的表达谱能够揭示很多它在体内的作用，也能帮助鉴定与其他基因的功能性联系。许多基因的表达限定于特定的细胞或发育结构，表明在这些位置基因具有特殊的功能。其他基因对外

界刺激有反应而表达。例如，它们可能在暴露于内源性信号（如生长因子）或者环境分子（如引起 DNA 损伤的化学物质）的细胞中开启或关闭。在这种情况下，我们不是无理由地假设，在某种程度上基因的功能参与了感知这样的信号或细胞对它们的反应。具有相似表达谱的基因可能参与相似的过程，从这一点上看，若一个孤独基因与某个有特征的基因具有相似的表达谱，则依据“连带罪责”可确定其功能。而且，突变一个基因可以影响其他基因的表达谱，帮助那些基因链接至功能性通路和网络。如果突变的基因编码一个转录因子，这将会是一种事实，因为那个转录因子异常的功能将影响在它控制下的所有基因。

转录物组分析为基因和其产物与功能性通路和网络相联系提供更多机会，并为药物开发提供了新的机遇

传统上，已经在逐个基因的基础上开展了表达分析，应用的方法（例如 Northern 印迹、RT-PCR、原位杂交等）被优化用于为个体基因研究（节 7.3）。一些技术经得起适度的多重反应，例如建立 10 或 20 个 PCR 相对容易，但是涉及几百个甚至几千个基因的高度平行分析是不可能的，因为需要大量的时间建立单独的反应，当然还有做这些实验的费用。尽管如此，全转录物组分析的优点是明显的。例如，我们可能希望在哮喘、多发性硬化症或炎症性肠炎中寻找改变了表达模式的基因。我们能够同时观察所有的基因，并因而鉴定表达谱改变的每一个基因，而不是选择几个表达可能受疾病影响的候选基因进行分析。这些基因中的一些可能已经是特征明确的，但其他可能是孤独基因，允许指定暂时性功能。这样的分析也有直接的实际效益。例如，在疾病状态明确上调的基因可能是有用的药物靶标，而下调的基因可能编码能够治疗性使用的蛋白质。基因表达整体分析的需求已有助于新技术的开发，允许同时监测几千个转录物，并比较不同样本中它们的丰度。两个主要平台是可用的，一个是在测序的基础上，包括从代表性 cDNA 群体中进行 DNA 序列抽样。这可认为是一个**开放系统**（open system），因为整个转录物组对于分析是开放的。另一个是在杂交的基础上，涉及 DNA 微阵列的使用。这可认为是一个**封闭系统**（closed system），因为仅有那些在芯片上出现的序列才能被测量。在考虑这些技术以及如何使用它们之前，有必要驱散一个普遍的神话。虽然转录物组分析方法通常描述成“转录分析”，重要的是认识到一个人并不是测量转录的速率，而是稳定状态的 mRNA 水平，它也同时考虑了 mRNA 降解的速率。

19.3.2　直接序列抽样是一个判定不同转录物相对丰度的统计方法

全基因表达谱能够通过 cDNA 文库抽样实现

也许研究转录物组最直接的方法是从 cDNA 文库中随机挑选克隆进行测序，并计数每个序列出现的次数。每个克隆的丰度代表了原始样品中相应转录物的丰度。如果足够的克隆被测序，那么统计分析能够提供一个基因相对表达水平的粗略估计（Audic and Claverie，1997）。这种方法已经用于鉴定差异表达的基因，但因为需要大规模测序，所以此方法还是艰巨的。一个可能的捷径是利用极短的称为**序列标识**（sequence signature）的序列样品，并同时读取它们中的许多序列。已经开发了几种技术来利用这

一序列标识（框 19.2），迄今为止最有影响的一个技术是基因表达系列分析（SAGE），这将在下面详细讨论。

框 19.2 基因表达整体分析的序列抽样技术

cDNA 文库随机抽样

测序随机挑选的克隆，并再次检索数据库以鉴定相应的基因。每个序列出现的频率为原始样本中不同 mRNA 相对表达水平提供了一个粗略的估计（Audic and Claverie，1997）。这是一项非常费力的方法，特别是当需要比较几个 cDNA 文库时。

EST 数据库分析

EST 是由随机 cDNA 克隆单程测序产生的标识。如果 EST 数据可用于一个既定的文库，则不同转录物的丰度可通过确定数据库中每一序列的出现来估计（例如，Vasmatzis *et al.*，1998）。这是一种快速的方法，优点是因为它完全可以利用"电子克隆"（in silico）方法来实施，但是它依赖于相关样品 EST 数据的可用性。

差异显示 PCR

这个方法是为快速鉴定两个或更多样品间差异表达的 cDNA 序列而发明的（Liang and Pardee，1992）。这个方法的分辨力不足以在一个反应中处理整个转录物组，所以标记的 cDNA 片段群体由 RT-PCR 产生，使用了一个有两个碱基突出的 oligo-dT 引物和一个任意引物。不同引物组合的使用产生了代表转录组亚组分的 cDNA 片段池。来自两个样品的等效扩增产物，即使用相同引物组合扩增的产物，随后在测序胶上一起电泳，条带密度在数量上的差异揭示了差异表达的 cDNA。这一技术适于追踪差异表达基因，但是假阳性比较常见，必须应用其他方法验证预测的表达谱。

基因表达系列分析（serial analysis of gene expression，SAGE）

在这个技术中，通过利用Ⅱ型限制酶的特殊性质，从许多 cDNA 中收集非常短的序列标识（**序列标签**，sequence tag）（正文和图 19.5）。标签连接在一起，形成长的多联体，再对这些多联体进行测序。通过计数某一特定标签出现次数计数来确定每个转录物的出现。尽管有技术上的要求，但 SAGE 要比标准的 cDNA 抽样更有效，因为每个测序反应可以计数 50～100 个标签（Velculescu *et al.*，1995）。

大规模平行标识测序（massively parallel signature sequencing，MPSS）

像 SAGE 一样，MPSS 技术利用Ⅱ型限制酶从许多 cDNA 中收集短序列标签。然而，与 SAGE（标签按系列克隆）不同，MPSS 依赖于一个流动池中附着于微珠的几千个 cDNA 的平行分析（Brenner *et al.*，2000）。该方法的原理是使用Ⅱ型限制酶在每个 cDNA 上暴露一个四碱基的突出端。有 16 种可能的四碱基序列，通过于一套 16 种不同的寡核苷酸接头杂交检测。每一个接头与一个不同的、标记了某一特定荧光标签的解码寡核苷酸杂交。接头含有一个Ⅱ型限制酶的位点，使得另外一个四碱基暴露，此过程重复进行。每轮的切割和杂交后，通过微粒成像，几千个 cDNA 序列就以四核苷酸块形式被读取。如同 SAGE 一样，每一序列被记录的次数能够用于确定相对的基因表达水平。

SAGE 涉及短序列标签的抽样，并将其连接在一起形成一个长的多联体

第一个开发的高通量序列抽样技术是**基因表达系列分析**或 SAGE（Velculescu *et al.*，1995）。它是建立在两个原理基础之上：

▶ 假设一个短核苷酸序列标签来源于一个转录物内某一确定位置，则它能够独一无二

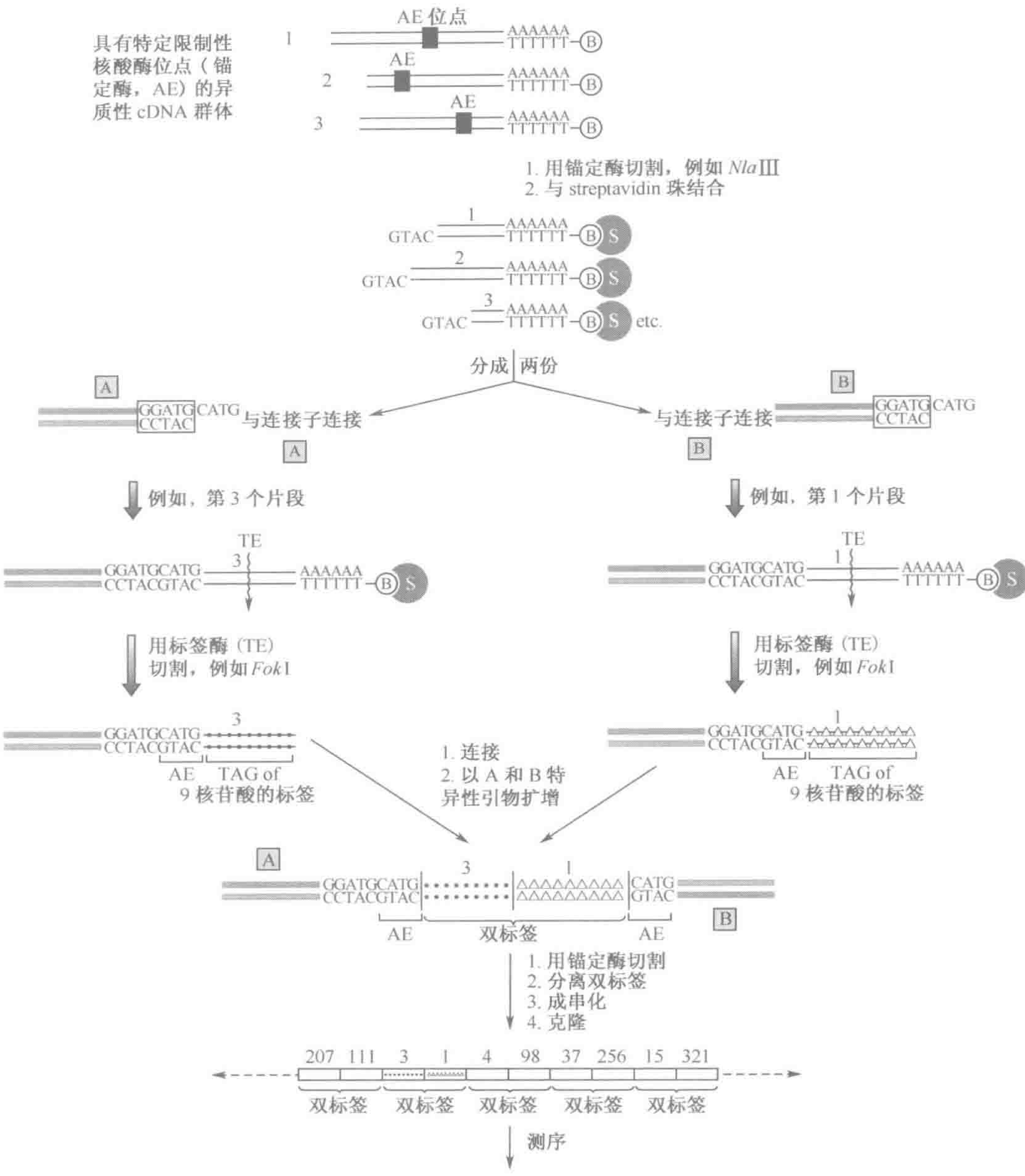


图 19.5　应用 SAGE 方法进行多重基因表达扫描

这个方法的基础是将每个 cDNA 分子分解成代表性的短序列标签（约 9 个核苷酸长）。然后，单独的标签连接在一起（多联体化）形成一单个长 DNA 克隆，见本图最底部所示。序列标签上的数字，代表衍生出标签的某一特定 cDNA。克隆测序提供了不同序列标签的信息，能够鉴定相应 mRNA 序列的存在。应用连接了一个生物素基团的一个 oligo（dT）引物，mRNA 被转换成 cDNA。同时生物素酰化 cDNA 被频繁切割位点的限制性核酸酶切割（锚定酶，AE；在本例中是 NlaⅢ，可在 4 碱基序列 CATG 的 G 之后迅速切割）。结果产生的 3′端片段含有一个生物素基团，接着通过与抗生物素蛋白链菌素（streptavidin）包被的珠子结合，选择性地被回收，并分离至两个池中，随后分别与两个双链寡核苷酸连接子 A 和 B 中的一个相连接。这两个连接子在序列上有差异，除了它们具有一个 3′CATG 突出端，以及紧随其后的一个常见的将作为标签酶（TE）的Ⅱ型限制性核酸酶的酶切位点。在本例中，它是 FokⅠ，识别序列 GGATG（以方框表示），但在下游 9/13 核苷酸处切割。经 FokⅠ切割，每个 mRNA 产生一个 9 bp 序列标签，来自独立池的片段混合在一起形成“双标签”，再按所示多联体化。

地鉴定出来自某一单独基因的此转录物。例如，虽然人类基因的总数大约 30 000 个，但一个仅有 9 bp 序列标签理论上能够分辨 4^9（262 144）个不同的转录物。实际上使用的是 12～15 bp 的序列标签，可以提供更强大的分辨能力。

▶ 短序列标签的多联体化允许以一个系列的方式有效地分析多个转录物。在一单个克隆内，来自不同转录物的标签能够共价地连接到一起，随后该克隆被测序，以鉴定克隆内的不同标签。一个测序反应可以分析最多 100 个不同的转录物。

为了实施一个 SAGE 分析，首先从待研究的资源提取 poly（A）＋ RNA，并使用一个生物素酰化的 oligo(dT) 引物，将其转换成 cDNA。后续操作在图 19.5 中描述。首先用限制酶切割 cDNA，如 *Nla* Ⅲ，因其具有 4-bp 的识别位点而可以频繁地切割（这称为**锚定酶**，anchoring enzyme）。释放的附带生物素的 3′端片段与抗生物素蛋白链菌素珠结合。抗生物素蛋白链菌素结合的 cDNA 被分成两份，这两份全部独立地与两个双链寡核苷酸接头 A 和 B 连接。每个接头的突出部分能够与锚定酶产生的突出配对，紧随其后的是**Ⅱ型限制酶**（type Ⅱs restriction enzyme）如 Fok Ⅰ 的 5-bp 识别位点。这类酶具有与众不同的、识别特定序列但在此序列外侧下游一定数量 bp 处切割 DNA 的特性。所以用这种所谓的**标签酶**（tagging enzyme）切割会产生一个短的标识序列，定义为 **SAGE 标签**（SAGE tag），与接头相连接。两个接头中的每一个也含有一个不同 PCR 引物的复性位点。在实验的下一阶段，把这两个接头-标签池混合并连接以便在一个接头的两侧形成**双标签**（ditag，两个 SAGE 标签连接在一起）。然后，使用 PCR 扩增这些分子。用原来的锚定酶切割扩增产物，释放双标签，纯化的双标签连接在一起并克隆。随后对克隆载体的插入片段进行测序，这使得标签按照系列的方式被读取。由标签出现的频率推导出相对的转录水平。

在初期的实验中，Velculescu 等（1995）报道了重获来源于胰腺 cDNA 的 840 个序列标签。在这些序列标签中，498 个代表了 77 种不同的转录物，最丰富的转录物均由已知的、具有胰腺功能的基因产生（如羧肽酶原 A1 出现 64 次而胰蛋白酶原 2 出现了 46 次）。SAGE 独特的优势在于数据是数字化的，所以很容易比较不同实验之间标签的频率，即使（实验）是在不同时间由不同实验室实施的。已建立了几个用于 SAGE 数据储存和比较的数据库。随着允许使用极少量起始材料的适应性变化，这种技术变得更加有用了（Velculescu and Volgelstein，2000）。

19.3.3 DNA 微阵列采用多重杂交实验以便同时检测几千个转录物的丰度

两种主要类型的 DNA 微阵列是以不同的方法制造的，但是每种设备表达分析的原理是相似的

DNA 微阵列（DNA microarray）是将不同的 DNA 序列固定成点阵的缩微模型装置。有两种主要的类型，一种是通过把 DNA 分子机械点样至一包被的玻片而制成的；一种是通过寡核苷酸原位合成产生的（Schena *et al*.，1995；Lockhardt *et al*.，1996：节 6.4.3）。两种装置均称为**微阵列**（microarray），但后者采用了更专业的名词——**高密度寡核苷酸芯片**（high density oligonucleotide chip）。有许多点阵芯片的商品化来源而

自制的设备可用于许多实验室。与此相比，高密度寡核苷酸芯片全部由美国的生物技术公司 Affymetrix Inc. 生产，作为**基因芯片**（GeneChip）销售。

尽管这两种设备采用完全不同的方法生产，但表达分析的原理大致相同（Harrington *et al*., 2000）。表达分析是建立在**多重杂交**（multiplex hybridization）的基础之上，使用一个标记 DNA 或 RNA 分子的复杂群体（图 19.6）。对于两种方法来说，某一特定来源的 mRNA 分子群体都需要全部反转录形成一代表性 cDNA 复杂群体。对于点阵芯片来说，反应混合物中掺入一个荧光偶联的核苷酸，所以 cDNA 群体被普遍地标记。对于基因芯片来说，未标记的 cDNA 通过生物素的掺入转换成一个标记 cRNA（互补 RNA）群体，随后被荧光偶联的抗生物素蛋白所检测。然后，标记的核酸复杂群体应用于芯片并杂交。芯片上每个单独要素或点含有 10^6～10^9 个相同 DNA 序列的拷贝，因此不太可能在杂交反应中完全饱和。在这些条件下，芯片上每个位置杂交信号的强度与混合物中那个特定 cDNA 或 cRNA 的相对丰度成比例。这依次地反映了原始资源群体中相应的 mRNA 的丰度。所以，能够在一个实验中监测几千个不同转录物的相对表达水平。

应用每种类型的设备进行表达分析的原理类似，但仍有一些重要的实际区别，反映了芯片（cDNA 或寡核苷酸）的特征性质和杂交反应的专一性。在点微阵列芯片中，每个要素代表一单个双链 cDNA，长几百个 bp，可由克隆文库或相似资源经 PCR 制备（图 19.6A）。至于寡核苷酸芯片，每个要素是一个短单链寡核苷酸，长 20～25nt，是在生产过程中合成在芯片上。因为寡核苷酸能够以任何一期望的序列组合成，包括那些储存在公共或私人数据库中的序列，故不需要维持一个克隆文库（图 19.6B）。此方法的一个优点是寡核苷酸序列可被挑选用于区分密切相关的转录物。对于 cDNA 芯片的要素，存在一个同源基因或选择性剪接变异体间交叉杂交的显著风险。然而，一个缺点是寡核苷酸芯片上选择包含的序列必须为已知的。相反，有可能从 cDNA 文库的未知克隆中生产 cDNA 微阵列。cDNA 微阵列的杂交特异性高，是由于代表每一要素的cDNA序列长度的原因（图 19.6C）。然而在寡核苷酸芯片中，杂交特异性低。所以，每个基因由 20 种不同的，沿着序列“步移”的寡核苷酸代表（图 19.6D）。有 20 种与靶序列**完全匹配的寡核苷酸**（perfect match 或 PM oligo）和 20 种含有一单个碱基**错配的寡核苷酸**（mismatch 或 MM oligo），以控制非特异性杂交。为了判断一特定基因的信号，将来自全部 20 个 PM 寡核苷酸的信号叠加在一起，而从总数中扣除来自全部 20 个 MM 寡核苷酸的信号。为使背景或非特异性杂交标准化，在 cDNA 微阵列上也包含了阴性对照要素。两种类型的设施也都含有阳性对照要素，通常用组成性表达基因如肌动蛋白代表。细菌基因可选择地包含在对照要素中，那么样品就能被适当数量的细菌 DNA 所“抑制”。

使用标记不同荧光的cDNA群体，通过芯片间比较或者与一单个芯片双重杂交能够证明样品间差异基因表达

在任何以阵列为基础的表达分析实验中，一个主要的技术障碍是不同阵列间数据的可重复性。如果必须进行比较分析（如鉴定疾病中上调的基因），这就特别麻烦，因为这种差异是由于实验的变异性，还是由于真正的基因表达改变造成的还不清楚。如果实验是在不同的实验室开展，使用的是不同设备生产的芯片，这个问题就复杂了。需要严谨的对照，使表达数据的交叉实验差异标准化。

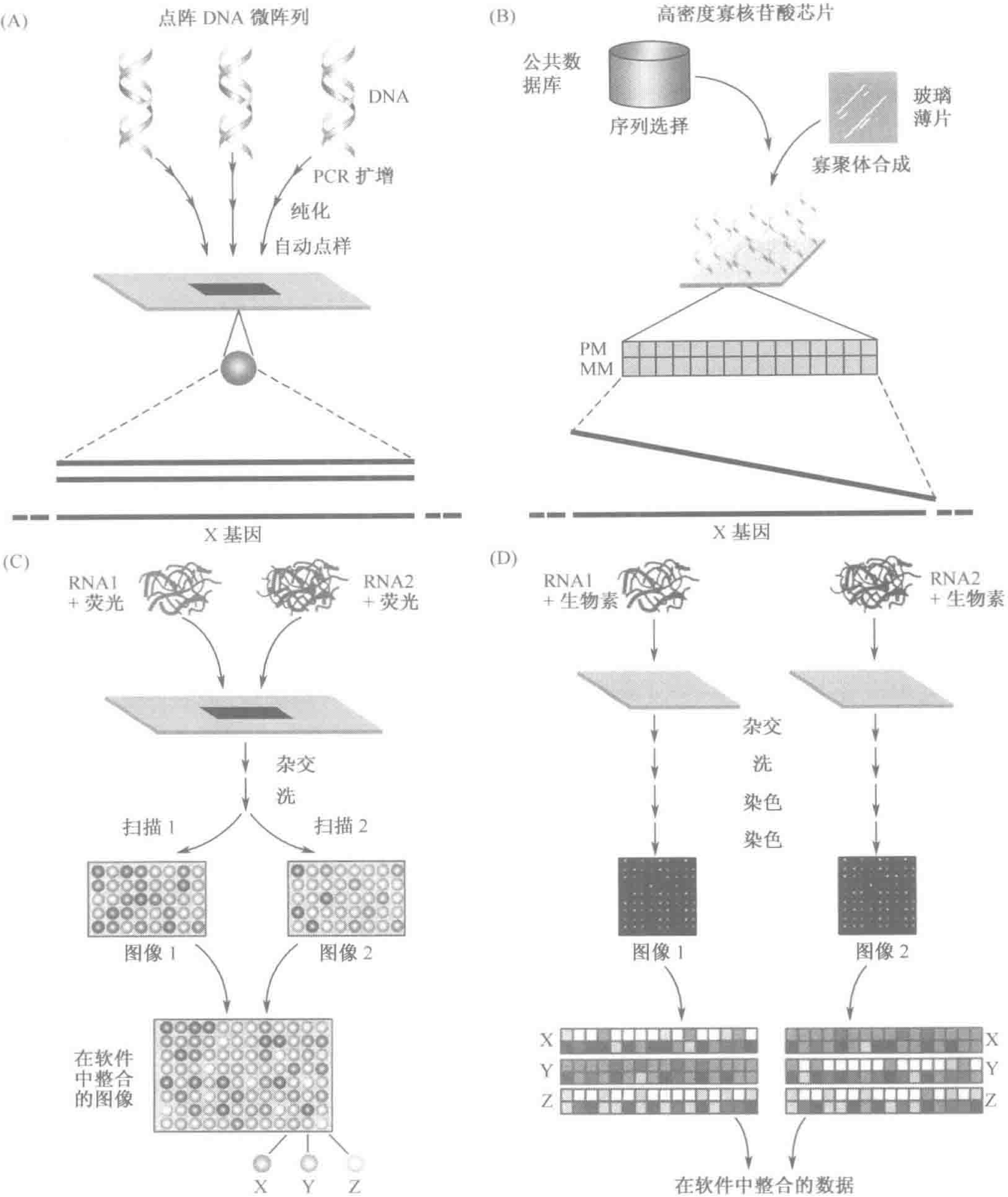


图 19.6　用 DNA 微阵列进行表达分析

(A) 扩增的 cDNA 分子经自动点样至玻片上生产点微阵列芯片。每个点或要素对应了一个几百 bp 或更多的连续性基因片段。预合成的寡核苷酸也能点在片子上（未显示）。(B) 高密度寡核苷酸芯片是利用光照指导结合性化学合成过程制造的，可在一块小玻璃芯片上以一个高度有序的排列方式产生几千个不同的序列。在芯片上，基因由 15～20 个不同的寡核苷酸对（PM，完全匹配；MM，错配）代表。(C) 在点阵芯片上，比较表达分析通常是用不同的荧光基团差异性标记两种 mRNA 或 cDNA 样品而进行的。它们与玻片上的要素杂交，随后独立地扫描检测两种荧光。图像底部标记为 x、y 和 z 的彩色点对应了在样品 1 中呈现升高水平（x），在样品 2 中水平升高（y），以及在样品 1 和 2 中水平相似（z）的假设基因。(D) 在 Affymetrix 的基因芯片，生物素酰化 cRNA 与芯片杂交并用偶联抗生物素蛋白的荧光染色。激光扫描检测信号。假设基因的成对寡核苷酸组显示在样品 1 呈现升高水平（x），在样品 2 种水平升高（y），以及样品 1 和 2 中水平相似（z）。经 Elsevier Science 允许从 Harrington 等（2000）修改。

能够避免上述问题的一个方法是用标记不同荧光的 cDNA 群体同时与同一芯片杂交。如上所讨论的，在不饱和的条件下，在每一要素处信号将代表样品中每个转录物的相对丰度。即使在一个芯片中，若使用了两个样品，然后样品经过完全性信号噪音比变异的标准化处理，则每个荧光基团信号的比值就提供了样品间一个直接的比较。在两种发射波长处扫描芯片，并用计算机合成图像，以伪彩色呈现图像。通常，一种荧光基团呈现绿色，另一种是红色。代表差异表达基因的要素呈现绿色或红色，而那些代表等量表达的基因呈现黄色（图 19.6c）。

19.3.4　DNA 芯片数据分析包括一个距离矩阵的产生和应用重复算法进行聚类有关数据点

在许多不同条件下基因表达数据分析能够显示基因间的功能联系

来自微阵列实验的原始数据是信号强度，必须进行标准化处理（校正背景影响和实验内变异，见上文）并检查由污染和极度远离中心值造成的误差（Yang *et al.*，2002）。将数据总结成一个标准化信号强度表格，表格的行代表单个基因而表格的列代表测量基因表达的不同条件。在最简单的例子中，表格有两列（如对照样品和疾病样品），这可以代表同时与芯片杂交的两个样品所产生信号的强度。然而，对能够使用的条件的数目并没有理论上的限制。例如，可能在一系列发育时间点或某一病毒感染发作后一系列时间点，或者当培养细胞暴露于一定范围的药物和其他化学物时检测基因表达。

分析的下一个步骤涉及具有相似表达谱的基因分组（Altman and Raychauhuri，2001；Quackenbush，2001）。通常，检测基因表达的条件越多，则分析越严格。图 19.7 显示了在三种条件下三个基因的分析。最初，基因 A 和 B 看起来好像功能相关，因为在条件 1 和 2 时它们的表达谱相似，而基因 C 看起来好像不同。然而，如果我们在分析时考虑条件 2 和 3 而忽略条件 1，那么在牺牲基因 B 的基础上，现在看起来好像基因 A 和 C 功能相关。许多条件下的比较有助于消除可能导致错误注释的假性关系。虽然一个简单的 3×3 矩阵分析能够用眼睛进行，但涉及几千个基因和几十种条件的表达数据，必须在计算机的辅助下进行提取。两类算法用于提取基因表达数据，一个是相似数据按等级聚类，一个是按照非等级方式确定聚类。

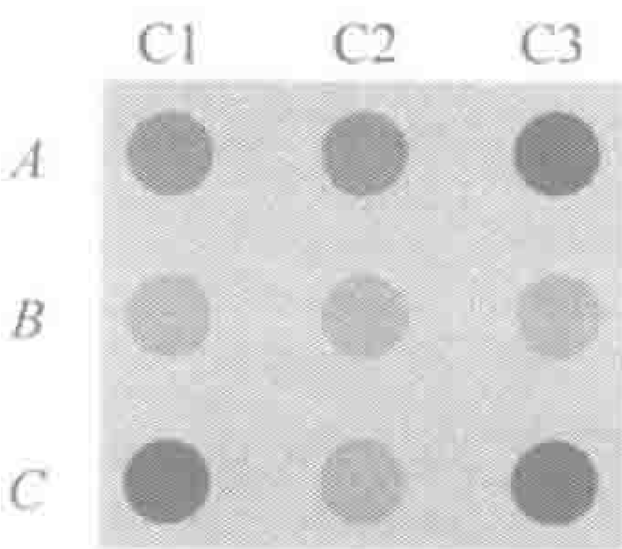


图 19.7　这个简单的例子显示在几种条件下监测基因表达分析的价值

在条件 1 和 2 下，基因 A 和 B 的表达谱看起来好像相似（即，它们都保持为变化小的），而基因 C 在条件 2 时看起来好像表达下调。这提示基因 A 和 B 是功能上相关的而 C 是不同的。可是，如果我们在条件 2 和 3 下观察表达谱而忽略条件 1，则看起来好像基因 A 和 C 是功能上相关的，而 B 是不同的。在全部三种条件下的分析提示任何基因之间没有关系。

等级聚类（hierarchical clustering）

等级聚类常用的方法是建立一个**距离矩阵**（distance matrix），列举芯片上每对要素间表达水平的差异。那些显示出最小区别的，表述为**距离函数**（distance function）d

的要素随后以级数的方式进行聚类。**凝聚式聚类**（agglomerative clustering）方法开始于对芯片上出现的每个基因分类作为单独聚类（即含有一个基因的一个簇）。搜索距离矩阵，具有最相似表达水平（最小距离函数）的两个基因，定义为邻居，随后被并入一单个的簇。重复此过程直至只剩下一个簇。为进一步的比较，如何计算合并的簇的表达数值存在着变异。

在**最近邻（单连锁）**[nearest-neighbor (single linkage)] 方法中，距离最小化。就是说，当两个基因 i 和 j 合并入一个单独的簇 ij，ij 和紧接的最近基因 k 之间的距离规定为两个数值 d（i，k）和 d（j，k）中较小的一个。在**平均连锁方法**（average linkage method）中，使用的是 d（i，k）和 d（j，k）的均值。在**最远邻（全连锁）**[farthest-neighbor (complete linkage)] 方法中，距离最大化。这些方法产生不同的树状图（图 19.8）。偶然地，代表芯片上所有基因的一个单独的簇可以应用分裂性聚类算法逐渐地裂解为独立的簇。



图 19.8 利用四个基因（A～D）的假设表达谱进行微阵列数据分析
等级聚类方法产生分支图谱（树状图，dendrogram），图中具有最相似表达谱的基因被聚集在一起，但是另一个聚类方法产生不同布局的分支图谱。左侧的模式是由最近邻（单连锁）聚类产生的典型布局，而右侧的模式是由最远邻（全连锁）聚类产生的典型布局。

非等级聚类（nonhierarchical clustering）

等级聚类的一个缺点是费时和资源紧张。作为一个替换方法，非等级聚类方法将表达数据分割为某一特定的预设数量的簇，因此大大地加速了分析，特别是当数据库非常大时。在 **k 均值聚类方法中**（k-means clustering method），分析开始时就确定了若干被称为**簇中心**（cluster center）的点，然后每个基因被分配至最适合的簇中心。依据每个簇的成员，重新计算均值，即重新定位簇中心。随后，重复分析，结果所有基因被分配至新的簇中心。这一过程反复重复，直至各种簇的成员不再变化。**自组织图谱**（self-organizing map）在概念上相似，但算法是从一神经网络的使用中提炼而来。

19.3.5 DNA 芯片已经用于研究人类细胞系、组织活检样品和动物疾病模型中整体基因表达

正常和疾病样品的比较能够用于鉴定疾病标记物和潜在的药物靶标

DNA 芯片已经广泛地应用于研究人类转录物组，也许这一技术最明显的实用性应用是疾病研究。存在于细胞系、活检组织或动物模型的正常和疾病组织的比较允许（我们）鉴定特异地表达于疾病状态或非疾病状态的基因。疾病特异的个体基因能够成为有

用的诊断标记。疾病特异的基因可能编码蛋白质，其存在或活性对疾病症状起作用，这些是治疗性干预的潜在性靶点。同样地，如果一个或多个基因的产物在疾病中缺如，这些蛋白质本身就可能是有用的治疗性因子。

McCaffrey 等（2000）的研究是上述方法一个提供信息的例子。这些研究者应用 Affymetrix 基因芯片来比较正常动脉和取自动脉粥样硬化损伤患者的动脉的表达谱。他们显示一个特殊的基因 *EGR1* 在疾病状态中上调 5 倍。这个基因编码一个转录因子，该转录因子控制编码生长因子和其他的信号分子，细胞黏附分子以及调节血液凝固的蛋白质的基因。这些下游蛋白在富含胆固醇细胞沉积于动脉内表面的过程中具有明显的作用，所以 *EGR1* 代表一个有用的药物靶标。

多重分析可以提供转录谱，用于区分非常相似的疾病和发现新的疾病亚型

单个标记对于所有疾病的诊断没有益处，特别是对那些密切相关的（例如，不同类型的癌症）。在这些病例中，芯片能用于 50 或 100 个提供较高分辨率的基因的转录谱。当疾病的种类是已知时，这一过程称为分类预测。（我们）期望实验结果符合某一特定数量的预定种类，正因为这个原因，数据分析描述为“监督式”。

这个方法的一个例子是应用芯片区分急性粒细胞性白血病（AML）和急性淋巴细胞性白血病（ALL）。虽然疾病相似，但它们对应不同治疗，因此正确的诊断是成功治疗所必需的。传统上，已使用联合实验技术，包括蛋白质标记的检测、差异染色、细胞遗传学分析和血涂片细胞的直观观察。没有实验是 100%准确的，单独的实验有时可能得出矛盾的结果。Golub 等（1999）使用大约有 7000 个人类基因的点阵芯片检查 38 个骨髓样品。应用如上讨论的自组织图谱算法，依据大约 50 个基因谱，他们正确地鉴定了每个样品是 AML 还是 ALL。

在某些时候，疾病样品的表达谱已经鉴定了先前未知的疾病亚型（分类发现）。因为在实验开始时并未确定种类，所以这种类型的分析描述为“非监督式”。例如，Alizadeh等（2000）应用 cDNA 芯片鉴定了非何杰金氏淋巴瘤的两个不同型（图 17.18）。只有 40%这类疾病的患者受益于药物治疗，直到现在，一直没有方法预测患者对治疗如何应答。然而，这两种新的疾病亚型看起来好像符合这个疾病的应答类型和非应答类型。相似地，Bitter 等（2000）能够鉴定皮肤黑色素瘤独特的两种亚型。

细胞系和动物模型能够用于在多种条件下或多个时间点检测基因表达谱

既然活组织检查允许我们并行比较健康和疾病间的基因表达谱，那么细胞系和动物模型则提供了检测不同条件或不同时间点的增长的多面性。例如，Zhu 等（1998）用巨细胞病毒感染培养的人类包皮成纤维细胞，并在感染后用含有 6000 个基因的 cDNA 微阵列分析几个时间点的基因表达谱。超过 250 个基因在感染过程中出现上调，包括几个已知的调整免疫反应的基因。细胞系也已用于研究生长因子和细胞因子的作用（如 Der *et al.*，1998）以及癌基因转染（如 Khan *et al.*，1999）。该方法的一个延伸是利用人类细胞系分析对药物、化学物和毒物的应答。这样的实验偶然地揭示了基因间无法预料的功能性联系。例如，Iyer 等（1999）研究了添加新鲜血浆时血浆饥饿细胞的转录物组，表明更新血浆后在不同时间点诱导出不同类型的基因。最先表达的基因是增殖反应基

因，如 *JUN* 和 *FOS*，但在随后的时间点，许多诱导的基因是那些涉及创伤愈合的基因，例如 *FGF7* 和 *VEGF*。

19.4　蛋白质组学

19.4.1　蛋白质组学包括蛋白质表达分析、蛋白质结构和蛋白质相互作用

蛋白质组分析显示与特定蛋白质的丰度改变如何协调细胞的生化活动

尽管转录物组分析显然对于基因功能特征研究非常有用，但不要忘记大多数基因的最终产物是蛋白质。蛋白质组学使得这些产物被直接研究，这是非常重要的，因为它暴露了转录物组分析的两个局限。

首先，部分归因于转录后调节，不是细胞中所有的 mRNA 都被翻译，所以转录物组可能包含在蛋白质组中找不到的基因产物。同样，在不同转录物间蛋白质合成和蛋白质更新的速度有差异，因此一个转录物的丰度未必与编码蛋白质的丰度一致（Gygi *et al*.，1999b）。因为这些原因，转录物组既不可能在质量上，也不可能在数量上精确地代表蛋白质组。

第二，蛋白质活性常常依赖于翻译后修饰，这在相应的转录物水平是不可预测的。许多蛋白质在细胞内作为无活性的分子存在，需要通过诸如蛋白水解切割或磷酸化过程才能激活。当一个特定的翻译后变异体的丰度的差异具有显著性时，这意味着只有蛋白质组学能够提供在基因表达和功能之间建立一个联系所需要的信息。例如，信号蛋白 stathmin 在各种癌症中高水平表达，包括儿童白血病，但仅有蛋白质的磷酸化形式是疾病的一个有用的标记。所以，为了充分了解细胞中存在的功能性分子，必须直接研究蛋白质组。

蛋白质结构和蛋白质相互作用分析能提供重要的功能信息

就像转录物组学，蛋白质组学能用于监测不同基因产物的丰度。我们能够在相关的样品间比较细胞中所有蛋白质的表达，鉴定具有相似表达模式的蛋白质，突出在蛋白质组中出现的重要变化，例如在疾病过程中或对特定外界刺激的应答过程中。这有时称为**表达蛋白质组学**（expression proteomics）。然而，蛋白质组学另一个重要的方面是常常通过研究**蛋白质相互作用**（protein interation）来确定功能。蛋白质在细胞中通过与其他分子间相互作用来进行它们的活动。因此确定蛋白质间特定的相互作用有助于确定蛋白质个体功能，并使蛋白质与通路和网络相联系。从蛋白质与小分子（可能作为配体、辅助因子、底物、构象调节因子等）以及核酸的相互作用中能够提取进一步的信息。蛋白质相互作用分析与蛋白质结构分析相互重叠。一个蛋白质三维结构的知识能够有助于预测与其他蛋白质和更小分子的相互作用，这对于有效药物的开发非常有益。蛋白质结构的比较也为确定基因间的进化关系以及研究它们的功能提供深层次的方法。

19.4.2　表达蛋白质组学的繁荣是通过两种主要技术平台的结合：二维凝胶电泳（2DGE）和质谱

蛋白质组含有成千上万个蛋白质，丰度相差四个或更多数量级。像核酸一样，蛋白

质能够通过特定的分子相互作用来检测和鉴定，在大多数情况中，使用抗体或其他配体作为探针。然而，与核酸不同，没有克隆或扩增罕见蛋白质的方法。而且，蛋白质的物理和化学特性多种多样，以至于没有单独的、通用的、类似于杂交的方法学，能够用于在一单个实验中研究整个蛋白质组。所以，当开发蛋白质芯片作为 DNA 微阵列直接的等价物具有明显的利益时（框 19.3），全蛋白质组分析普遍需要替代性的技术平台。

框 19.3　蛋白质芯片

DNA 微阵列能够用于全基因组分析，因为所有 DNA 分子在化学上相似，并能在一个实验中杂交。相反，蛋白质在化学上变化很大。一些是酸性，而其他是碱性，一些是带电的或有极性，而其他是疏水的。许多蛋白质经历进一步改变它们化学和物理性质的翻译后修饰。因为这个原因，真正的“通用”蛋白质芯片是难以生产的。可是，在过去几年中，蛋白质芯片技术已经有了一些技术上的进步，达到顶峰的是一个酿酒酵母全蛋白质组芯片的创造（Zhu *et al.*，2001）。已有各种不同类型的蛋白质芯片的描述（Zhu and Snyder，2003）。

- ▶ **抗体芯片**（antibody chip）：这些由排成阵列的抗体组成，用于在一复杂混合物中检测和定量特定蛋白质。它们被认为是小型化的高通量免疫试验设备；
- ▶ **抗原芯片**（antigen chip）：抗体芯片的反转，这些设施包括排成阵列的蛋白质抗原，用于在一复杂混合物中检测和定量抗体；
- ▶ **通用蛋白质芯片（功能性阵列）**[universal protein chip（functional array）]：这些设施包括在表面排成阵列的任何种类的蛋白质，能用于检测和定性特定蛋白质-蛋白质和蛋白质-配体相互作用。能够采用各种检测方法，包括在溶液中标记蛋白质或检测芯片的表面特性变化，如通过表面胞质共振。**植物血凝素芯片**（lectin chip）包括在这一类别内，可用于检测和定性糖蛋白；
- ▶ **蛋白质捕获芯片**（protein capture chip）：这些设施不包含排成阵列的蛋白质，而是其他作为广谱的或专一性捕获因子与蛋白质相互作用的分子。例子包括含有作为特定捕获因子的分子印记多聚体的寡核苷酸类似物和芯片，或者由像 BIAcore Inc. 和 Ciphergen Biosystems Inc. 公司生产的专有蛋白质芯片，采用基于表面化学差异的广谱捕获因子，以简化复杂蛋白质混合物。
- ▶ **溶液阵列**（solution array）：最新一代蛋白质芯片，来源于二维阵列形式，以增加它们的适应性和操控能力。例如，这样的设施可以建立在编码的微珠和条码化的金纳米颗粒基础之上。

现在，表达蛋白质组学是建立在“分离和展示”技术基础之上，即复杂的蛋白质混合物被分离为它们的组成成分，以便获取感兴趣的要素（例如，在一个疾病样品中存在但在匹配的健康样品中缺如的蛋白质）用于进一步的特征研究。分离通常采用**二维凝胶电泳**（two-dimensional gel electrophoresis，2DGE）进行，这是一种发展超过 25 年的技术，具有在一单块胶上最多分辨 10 000 种蛋白质的能力（Gorg *et al.*，2000；Herbert *et al.*，2001）。直到几年前，蛋白质组学的主要瓶颈仍是由未知的“点”代表的分离蛋白质的特征研究。如上讨论的，蛋白质鉴定和定量的一个方法是应用抗体或其他特异性探针。然而，这种方法仅能应用于任何一个实验中的少量蛋白质，这当然是由这类探针的可用性决定的。伴随注释技术发展而来的突破建立于质谱基础之上，质谱能够应用于任何蛋白质，并能大规模完成（Griffin *et al.*，2001；Mann *et al.*，2001）。这需要设备设计的创新，以及新型生物信息学方法用于使用肽质量数据进行数据库检索（Aebersold and Mann，2003）。

2DGE 是蛋白质组学中蛋白质分离的主要平台，但并非没有限制

有许多不同的蛋白质分离方法，而所有的方法都是利用蛋白质与蛋白质之间不同的特殊化学或物理性质，如质量、大小、电荷、可溶性和与不同配体的亲和力。就像可以预期的那样，在任何分离技术能够利用的性质越多，分辨能力就越大。2DGE 的原理是依据蛋白质的电荷和质量来分离蛋白质，每次分离发生于一个不同的维度（Gorg *et al.*，2000；Herbert *et al.*，2001）。一个复杂的蛋白质样品上样于一变性聚丙烯酰胺凝胶，第一相以**等电聚焦**（isoelectric focusing）分离。在这个技术中，蛋白质按 pH 梯度迁移，直到它们到达其**等电点**（isoelectric point，考虑局部 pH 时它们的电荷为中性的位置）。标准程序是制备一**固定** pH **梯度凝胶**（immobilized pH gradient，IPG gel），在此胶中，缓冲基团与聚丙烯酰胺基质相结合，这可以预防它们在长胶泳动过程中漂移和变得不稳定。然后凝胶在去污剂十二烷基磺酸钠中平衡，它按化学当量与变性的蛋白质骨架结合，使其带上大量负电荷，有效地去除个体蛋白质间任何电荷的差异。所以第二相的分离是依赖于蛋白质的质量，越小的蛋白质在穿越胶孔的时候，移动得越快。然后，凝胶染色，蛋白质以点的形式显现出来（图 19.9）。

虽然 2DGE 具有高分辨率，并且是蛋白质组学中应用最广泛的蛋白质分离技术，但在代表性、敏感性、可重复性和便利性（特别是在自动化适宜性方面）方面存在几个其使用的限制因素。几类蛋白质在标准凝胶上是低显现的，包括：非常碱性的蛋白质、

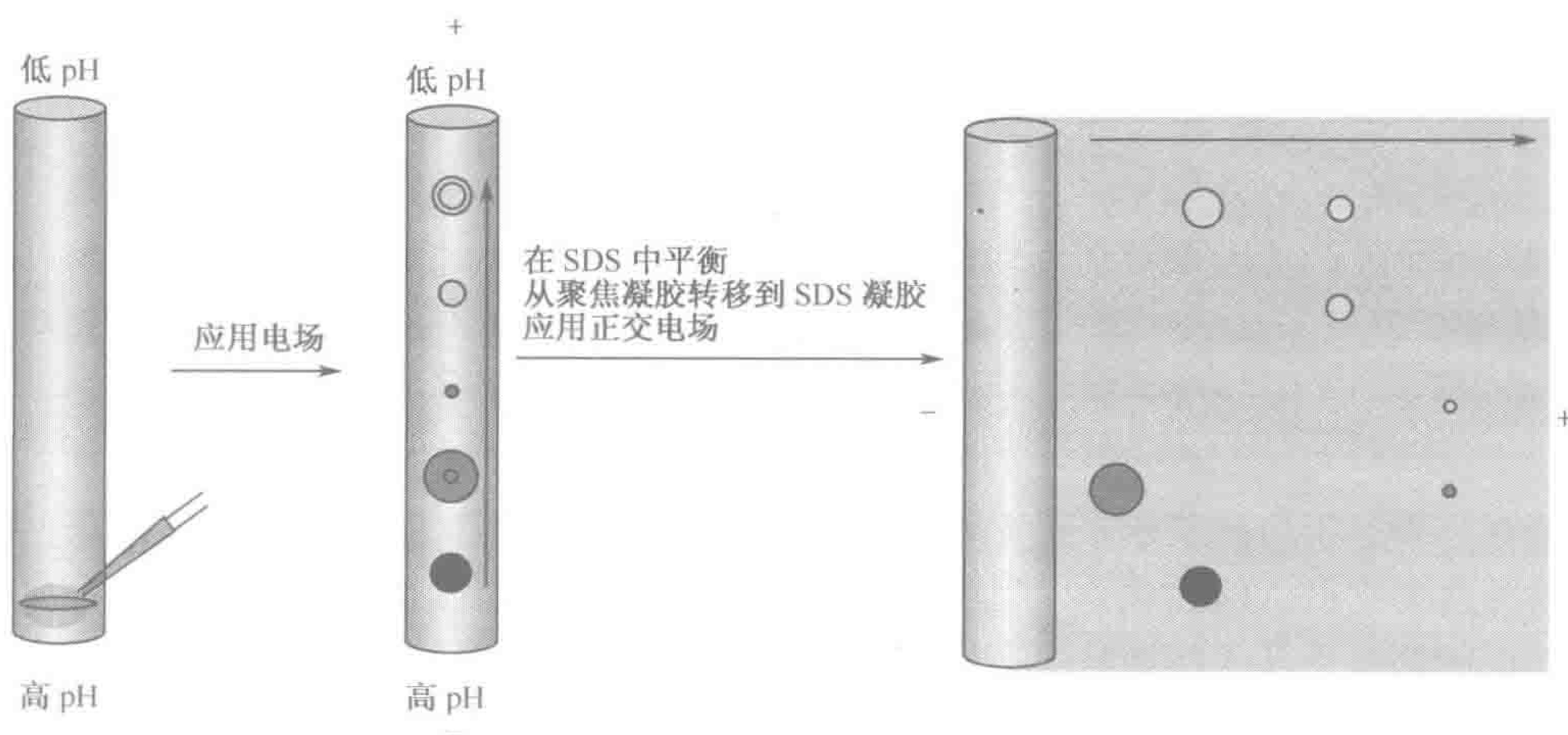


图 19.9　二维凝胶电泳的原理

一个复杂的蛋白质混合物上样于一等电聚焦凝胶的碱性末端（这常常是一个管状的凝胶或预制的胶条，由固定 pH 梯度的聚丙烯酰胺基质构成）。凝胶在电场中处理，蛋白质向它们的等电点迁移，在等电点处 pI 值与周围的 pH 相等，净电荷为零。这依据蛋白质的电荷（酸性蛋白质以黑色表示，碱性蛋白质以浅灰色表示，深灰色的阴影表示中等 pI 值的蛋白质）分离蛋白质。虽然凝胶依据它们的大小过滤蛋白质，但长胶泳动确保所有蛋白质到达它们的等电点，并且系统达到平衡。然后，聚焦凝胶在 SDS 中平衡（SDS 以恒定的质量比与所有变性的蛋白质结合），连接至一个标准的 SDS-聚丙烯酰胺凝胶。随后蛋白质按大小分离（由不同直径的圆圈表示），经过凝胶时，较小的蛋白质要比较大的迁移更远。

在水性缓冲液中溶解度差的蛋白质以及膜蛋白。为了有效分离蛋白质，可能必须预分馏蛋白质，并使用不同的去污剂和缓冲液，以提取不同类型的蛋白质。2DGE 的敏感性依赖于对非常稀有蛋白质的检测极限，但这个问题由于名为 SYPRO 染料的非常敏感的染色试剂的开发而已被解决，SYPRO 染料可以检测纳克范围的蛋白质点。敏感性也受到凝胶分辨率的影响，因为当那些代表高丰度蛋白质的点使代表稀有蛋白质的点模糊时，就难以检测代表稀有蛋白质的点。这些问题能通过预分馏，去除高丰度蛋白，简化上样凝胶的初始样品，提高分离的分辨率来解决。在后面的例子中，通过使用非常大的凝胶能够提高分离距离，但一个更方便的替代方法是使用**放大凝胶**（zoom gel）即具有非常窄 pH 范围的凝胶进行等电聚焦（图 19.10）。对于全蛋白质组分析来说，使用一台计算机能够将放大凝胶获得的图像接合在一起。

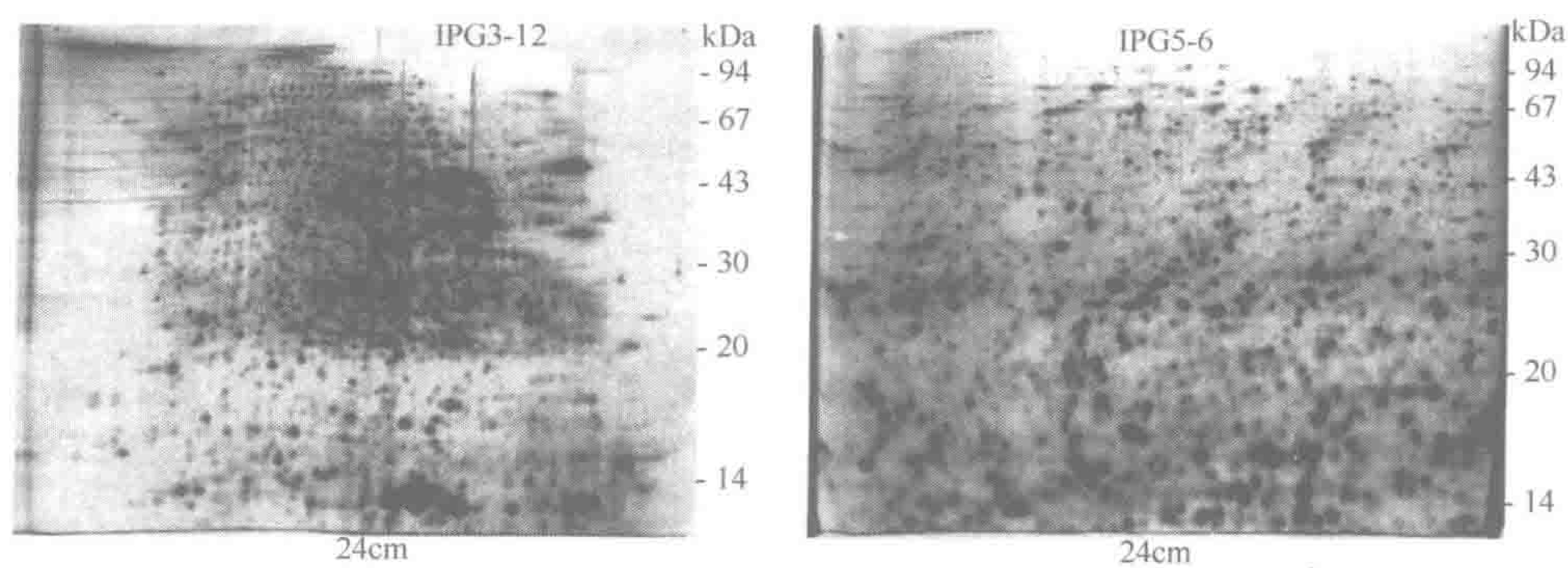


图 19.10 窄 pH 范围凝胶的分辨能力

两个图像都显示了二维凝胶电泳分离并银染的小鼠肝脏蛋白质，以揭示单个的蛋白质点。上部的图像是一个宽 pH 范围凝胶（pH 3～12），而底部的图像是一窄 pH 范围凝胶，在 pH 5～6 范围放大蛋白质点。注意在较宽范围的凝胶中，大多数蛋白质聚集在中间，反映出大多数蛋白质拥有范围 4～7 的 pI 值的事实。再引自 Orengo 等（2002）．Bioinformatics．BIOS Scientific Publishers 出版。

2DGE 的一个主要局限就是不适合于高度自动化，而这是许多样品高通量分析所必需的。有必要开发点识别和定量的软件，以及能够比较多个凝胶间蛋白质点的算法和自动挑取感兴趣的点并加工它们用于 MS 的自动化装置。2DGE 遇到的许多困难可由开发基于多维高压液相色谱（HPLC）的替代性分离方法而解决。这些方法更快、更敏感，不需蛋白质染色。它们保证更精确的定量，能分辨 2DGE 中低显现的蛋白质，且更易于自动化以及与下游分析相整合。虽然缺乏染色凝胶的可视性，但液相色谱方法可能最终替代 2DGE，成为蛋白质组学中蛋白质分离的主要平台（Lesney *et al.*，2001；Wang and Hanash *et al.*，2003）。

质谱是用于未知蛋白质高通量注释的唯一通用方法

质谱（mass spectrometry，MS）用于确定一特定样品或分析物中精确的分子质量。MS 进行蛋白质注释的原理是使用精确分子质量作为查询对象，检索数据库（Mann *et al.*，2001；Aebersold and Mann，2003）。该方法能比另一种注释方法——Edman 降解法——进行蛋白质直接测序，完成得更快，并且易于自动化进行高通量样品分析。当一个

人考虑有必要处理上千个来自 2D 凝胶的点或几百个 HPLC 组分时，这就变得重要了。

直到最近，MS 才应用于如蛋白质和核酸的大分子，因为它们在电离的过程中被打断成随机片段。现在**软电离**（soft-ionization）方法，如基质辅助激光解吸附/电离（MALDI）和电喷雾电离（ESI）（框 19.4），容许这样的分子电离却不会片段化（图 19.11）。通过使实验确定的质量与数据库预测的质量相关联，蛋白质质量和更常用的经蛋白酶消化从蛋白质得到的肽片段的质量，能够用于鉴定蛋白质。通过 MS 进行蛋白质注释有三种不同的方法（图 19.12）。

框 19.4　蛋白质组学中的质谱

质谱仪（mass spectrometer）

一台质谱仪具有三个部分。**离子发生器**（ionizer）将分析物转换为**气相离子**（gas phase ion），并使它们向**质量分析器**（mass analyzer）加速，质量分析器根据离子的**质量/电荷比**（mass/charge ratio）在它们去向离子检测器的途中将其分开，离子检测器可以记录单个离子的影响，并把这些作为分析物的**质谱**（mass spectrum）显示出来。

软电离方法（soft-ionization method）

大分子没有片段化和降解的电离被称为**软电离**（soft-ionization）。蛋白质组学中广泛使用的是两种软电离方法。**基质辅助激光解吸附/电离**（matrix-assisted laser desorption/ionization，MALDI）涉及在一有机溶剂中将分析物（来源于某一特定蛋白质样品的胰蛋白酶消化肽）与一光吸收基质复合物混合。溶剂气化产生分析物/基质结晶体，被一短脉冲激光能量加热。激光能量以热的形式解吸附，导致基质和分析物膨胀，进入气相。随后分析物被电离并加速至检测器。在**电喷雾电离**（electrospray ionization，ESI）中，分析物溶解，溶液经一根狭窄的毛细管推入。穿越孔径的电位差导致了分析物以带电粒子的一个纤细喷雾形式出现。当离子进入质量分析器时液滴气化。

质量分析器（mass analyzer）

用于蛋白质组学的两个最简单类型的质量分析器是**四极杆**（quadrupole）和**飞行时间**（time of flight，TOF）分析器。一个四极杆分析器由四根金属棒组成，每对金属棒用电力连接，带有相反的电压，电压可由操纵者控制。通过变化加在穿越离子束上的电位差获得质谱，使得不同质量/电荷比的离子对准检测器。飞行时间分析器测量离子从飞行管运行到检测器所花费的时间，是一个由质量/电荷比所决定的因素。

串联质谱（tandem mass spectrometry）

这涉及两个或更多的质量分析器的系列使用。已有各种各样的 MS/MS 设备，包括三个四极杆、混合的四极杆/飞行时间设备。质量分析器由一个**碰撞室**（collision cell）分隔开来，而碰撞室含有惰性气体，可使离子分裂成片段。第一个分析器选择某一特定肽离子，并使其直接进入碰撞室，在此处它被片段化。随后片段质谱被第二个分析器获取。这两种功能可能在更高级的设备中兼有，如**离子陷阱**（ion trap）和**傅里叶变换离子回旋加速器**（Fourier transform ion cyclotron）。

- **肽质量指纹图**（peptide mass fingerprint，PMF）：一种简单的蛋白质混合物（典型地是来源于 2D 凝胶的一单个点）由胰蛋白酶消化产生一个胰蛋白酶消化肽的集合。这些有待于 MALDI-MS 飞行时间（TOF）分析（框 19.4），可以得到一套质量光谱。这些光谱作为一个检索查询序列用于检索查询 SWISS-PROT 数据库。检索算法贯彻数据库中所有蛋白质的有效胰蛋白酶消化物，并计算预测胰蛋白酶消化肽的质

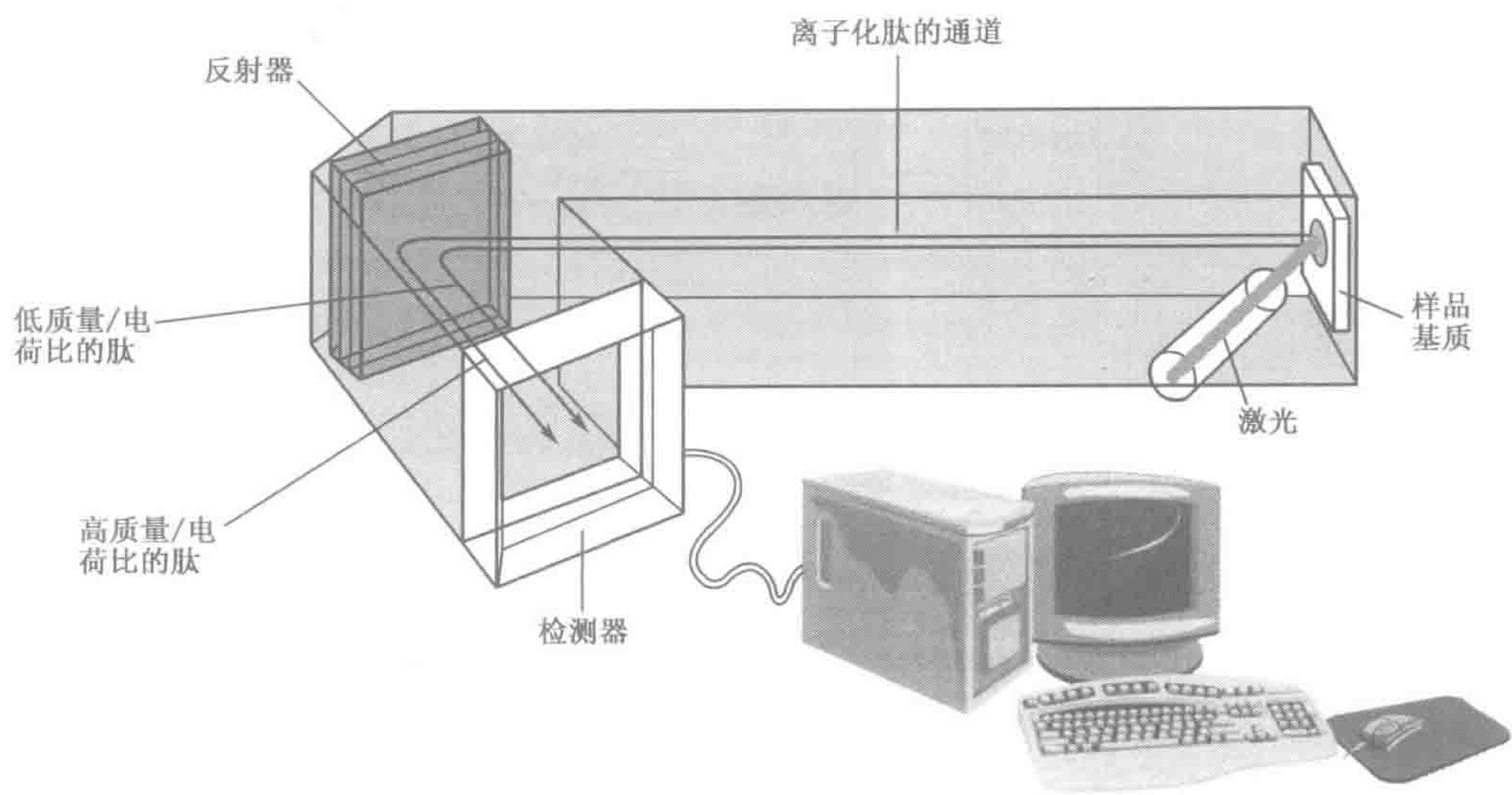


图 19.11　MALDI-TOF MS 的原理

分析物（通常是一个胰蛋白酶肽片段的集合）与一基质复合物混合，放置于激光光源的附近。激光加热分析物/基质结晶体，导致分析物膨胀为气相且无明显的片段化。随后离子从一个激光管运行降至一个反射器，它可将离子聚焦到一个检测器。飞行时间（离子到达检测器所用的时间）由电荷/质量比决定，使得分析物中每个分子的质量记录下来。

量。然后，尝试将这些预测质量与实验性确定的质量匹配。

- **片段离子检索**（fragment ion search）：如果 PMF 失败，就采用这个方法。胰蛋白酶消化肽片段经串联质谱（MS/MS；框 19.4）分析，在这一过程中肽被打断成随机片段。源自这些片段的质谱能用于检索 EST 数据库，但这个 EST 数据库不能检索 PMF 数据，因为完整的肽通常太大。随后，任何 EST 匹配序列能够用于 BLAST 检索，以鉴定推测的全长同源序列。一个称为 MS-BLAST 的专用算法适用于处理由肽片段离子得到的短序列标识。
- **肽阶梯的从头测序**（De novo sequencing of peptide ladder）：在这个技术中，由 MS/MS 产生的肽片段被排列成仅一个氨基酸长度不同的嵌套形式。通过比较这些片段的质量和氨基酸的标准表格，就有可能从头推导肽片段序列，即使无法在数据库中得到一精确的匹配序列。实际上，从头测序方法因为两个片段系列的存在变得复杂：即一个位于 N 端的嵌套和一个位于 C 端的嵌套。这两个系列能通过附着于蛋白质任一末端的诊断性质量标签来区分。

在一般的方法中，PMF 是最先尝试的，如果不成功，就需尝试其他不太精确的方法。PMF 最适合于简单蛋白质组的分析，诸如酵母蛋白质组，几乎没有剪接变异体和翻译后修饰。片段离子分析更适于复杂蛋白质组的分析，算法经过修改后，考虑了已知的翻译后修饰的质量。然而，不可能在序列的水平（如多态性）或在蛋白质修饰的水平（如复杂多糖）考虑所有的变异体。在这样的情况下，从头测序可以提供序列标识，能作为检索查询序列用于鉴定数据库中的同源性序列。

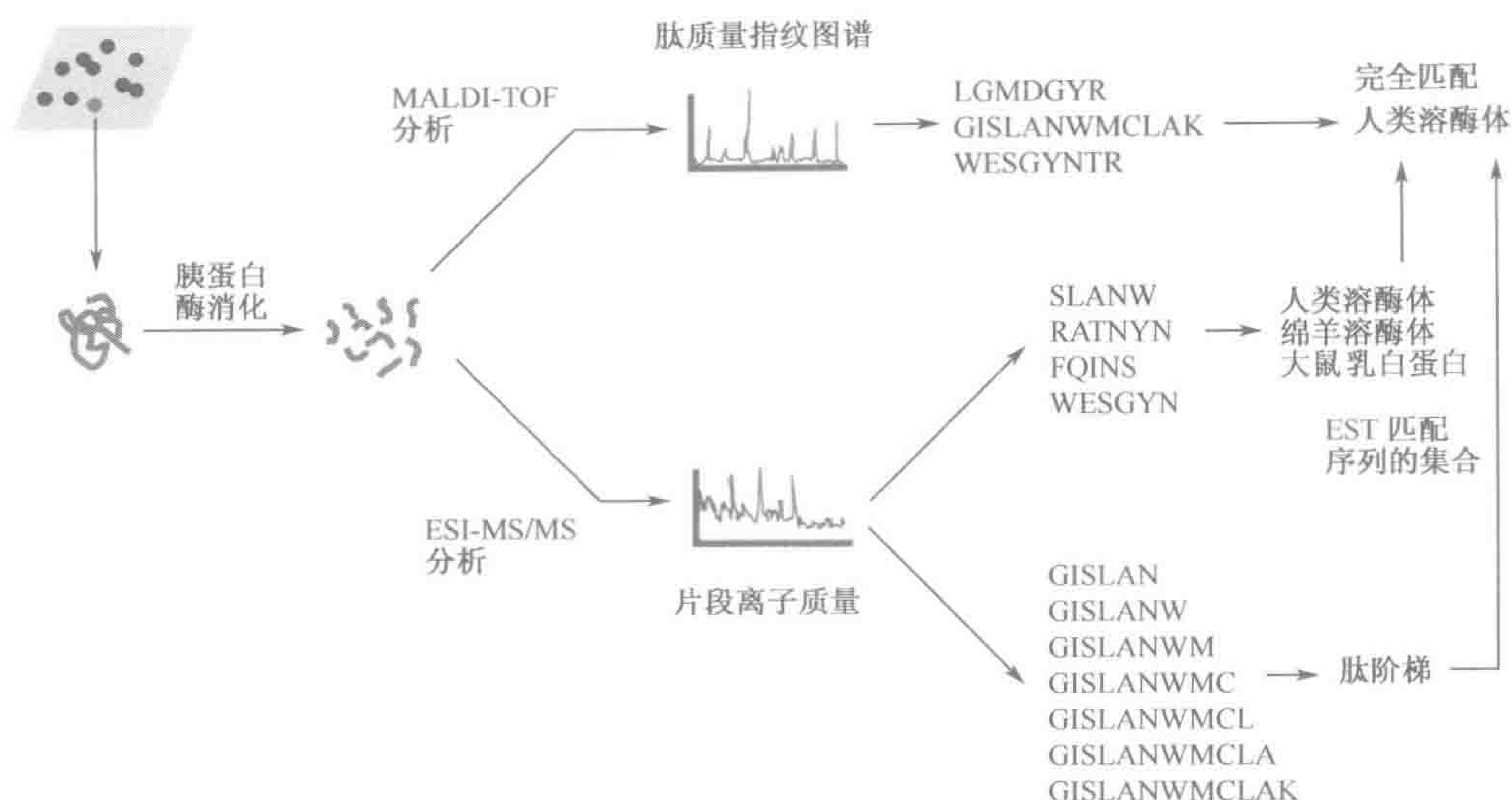


图 19.12　用质谱方法进行蛋白质注释

单独的蛋白质样品（如来自 2D 凝胶的点）被胰蛋白酶消化，只要临近的残基不是脯氨酸，就在赖氨酸（K）或精氨酸（R）残基的 C 端侧切割。胰蛋白酶消化的肽能作为完整的分子通过 MALDI-TOF 进行分析，质量用作查询目标检索蛋白质数据库。使用的算法是获得蛋白质序列，以相同酶如胰蛋白酶特异性切割，并比较这些肽的理论质量和通过 MS 得到的实验质量。理想地，几个肽的质量会确定同一个亲本蛋白，本例中是人类溶菌酶。如果蛋白质不在数据库中，可能没有匹配序列，或者更可能是，在实验过程中蛋白质已经受到翻译后修饰或人工修饰。在这些情况下，ESI-串联质谱能用于使离子片段化。片段离子质量能用于检索 EST 数据库，获得部分匹配，这将引导最终的正确注释。其他可选择的，肽阶梯质量能够用于从头确定蛋白质序列。

MS 也能用于分析差异表达蛋白

许多蛋白质组学实验的目的是鉴定在两个或更多样品中丰度明显不同的蛋白质。能够达到此目的的一个方法是检查 2D 凝胶，鉴定出现数量性变异的点，有几个软件包可用于辅助进行此类比较性研究。一个可供选择的方法是将来源于两个不同样品的蛋白质通过偶联标记上 Cy3 和 Cy5，并在同一块凝胶上分离它们（Patten and Beecham，2001；Rabilloud，2002；图 19.13）。这个方法称为**差异凝胶电泳**（difference gel electrophoresis，DIGE），利用与使用 DNA 微阵列进行差异基因表达（如节 19.3.3 所讨论的）同样的原理。而另一个解决问题的方法是用**同位素编码亲和标签**（isotope coded affinity tag，ICAT：Gygi *et al.*，1999a；Sechi and Oda，2003）标记不同来源的蛋白质。质谱仪能容易地区分和定量化学上相同并能共同纯化的同一复合物的两个同位素标记形式。所以，已经开发了利用生物素酰化的碘乙酰胺衍生物，在半胱氨酸残基选择性标记蛋白质混合物的 ICAT 方法。在胰蛋白酶水解后，生物素标签使我们可以对有标签半胱氨酸的肽进行亲和纯化。ICAT 试剂有“重”（d8）和“轻”（d0）同位素标记形式可用，这可以用于在不同条件下（如健康和疾病）差异性标记细胞池。标记后，合并细胞、裂解并分离蛋白质，为的是发生在两个样品中纯化损失也相同。当肽进入质谱仪时，比较它们的同位

素强度。如果它们是相等的，那么就没有下调或上调发生，则这个蛋白质就没有直接的重要性。如果强度不同，那么就发生了蛋白质表达的变化，则这个蛋白质就很重要。测量两种形式的数量，d0 形式的肽片段化并经数据库检索鉴定，如正文所讨论。

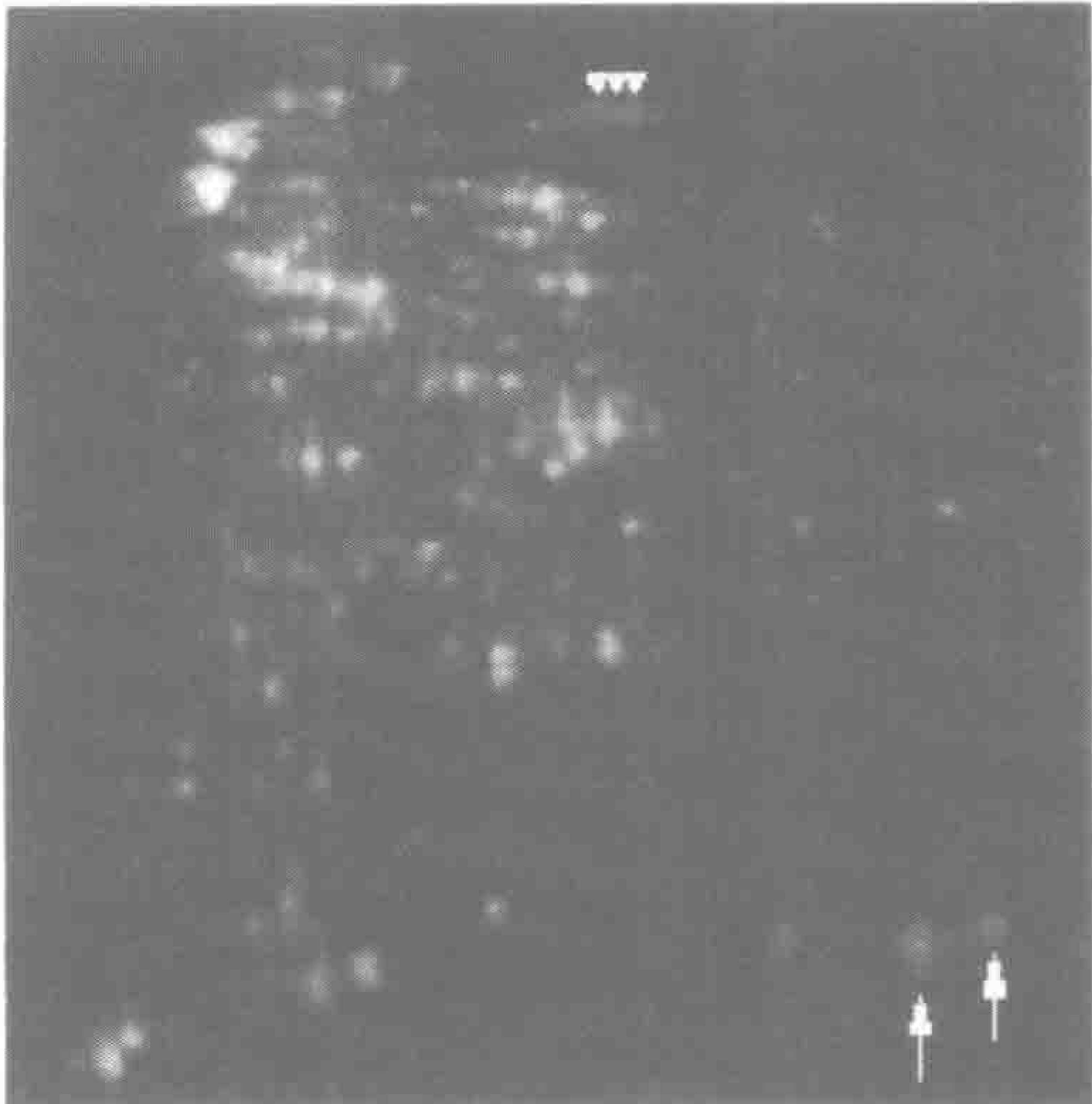


图 19.13　差异凝胶电泳原理的示范

来自细菌胡萝卜软腐欧文氏菌（*Erwinia carotovora*）的相同蛋白质样品用 Cy3 和 Cy5 标记，但 Cy5 样品中的两个待测蛋白——肌球蛋白和伴白蛋白是正常浓度的八倍。样品中丰度相同的蛋白质呈现黄色，因为 Cy3 和 Cy5 信号在图像叠加时强度相等，三个伴白蛋白异构体和两个肌球蛋白异构体，在 Cy5 标记群体中丰度更高，呈现为红点。经 Elsevier Science 允许再引自 Lilley 等（2001）。

19.4.3　表达蛋白质组学已经应用于研究与疾病和毒理相关的蛋白质改变

蛋白质组学能揭示转录物组分析无法鉴定的疾病标记和潜在的药物靶标

在健康和疾病样品中，或者在代表某一疾病进展的样品中，特定蛋白质的不同丰度能够有助于发现有用的标记和新的药物靶标。例如，乳腺癌中已经鉴定出几个表达水平异常的蛋白质，包括 PNCA（增殖细胞核抗原，DNA 聚合酶的一种成分）和各种热休克蛋白（Franzen *et al.*，1996）。在结直肠癌中，也显示热休克蛋白上调，但发现环氧化酶 2 和一个脂肪酸结合蛋白显著下调（Stulik *et al.*，1999）。当膀胱癌从早期的过渡期上皮进展为成熟的鳞状细胞癌时，各种标记包括不同类型的角质蛋白均表达。这些可用作评估分化程度的标记，因而监测疾病的进展。但是，注意处理样品时要小心，因为从头发和皮肤的 2D 凝胶点切取的微量蛋白质易于污染，并且它们也富含角质蛋白。

虽然表达蛋白质组学和转录物组学的应用相似，但是蛋白质组学具有优点，它从细胞中真正的功能分子取样，并考虑翻译后修饰。我们已经将 stathmin 作为一个实例进行讨论——这个信号蛋白在儿童白血病中上调，但只有磷酸化形式与疾病相关。蛋白质组学也有利于不含有 mRNA 的体液的分析。例如，Celis 等（2000）已发现一个称为 psoriasin 的蛋白质在膀胱癌患者的尿中含量丰富，能用作疾病早期诊断的标记。

毒物蛋白质组学有助于确定药物不良反应的基础

如在第 21 章讨论的，由于药物受体，以及决定药物如何吸收、代谢、分泌的酶和转运体蛋白的多态性变异，个体对药物有不同的反应。通过显示药物治疗后蛋白质组的变化，蛋白质组学能在药物不良反应预测和研究中发挥重要的作用。因为药物是与蛋白质直接相互作用并改变其行为的分子，所以许多药物反应不影响 mRNA 的丰度进而对转录物组没有影响。

作为一个例子，我们考虑一下免疫抑制性药物**环孢菌素**（cyclosporin）A。这广泛用于阻止移植或器官移植后排斥，特别是在儿童。环孢菌素 A 的一个主要副作用是肾毒性，发生于近 40%的患者。毒性与尿中的钙流失及其造成的肾小管钙化有关。大鼠，以及接下来的人类在未治疗患者和那些使用环孢菌素 A 治疗患者的肾脏蛋白质组学分析显示，一个特定蛋白质——**钙结合蛋白**（calbindin）水平显著不同（Aaicher *et al*.，1998）。这个蛋白在环孢菌素 A 治疗的人和大鼠的肾脏中丰度更低，直接表明环孢菌素 A 肾毒性的机制。有趣的是，钙结合蛋白在环孢菌素 A 治疗的猴子中没有损失，因而它们也没有遭受与人类相同的药物不良反应效应。因此研究环孢菌素 A 在猴子体内的代谢途径可能使我们获得避免人类毒性的机制。

19.4.4 蛋白质结构提供重要的功能信息

甚至当序列变异到无法再识别某一同源关系的程度，蛋白质结构也可能是保守的

在本章伊始，我们显示相似的蛋白质序列通常具有相似的结构，因而具有相似的功能。一个蛋白质的功能（或其中一个结构域）是由它的四级结构，也称为**折叠**（fold）所决定的。这形成了真正执行蛋白质生化活性的结合位点、相互作用的结构域和催化口袋。在这些结构中，少数氨基酸残基对特定的化学反应绝对关键，如一个酶与底物作用的活性部位内的残基。可是，一个蛋白质中大多数氨基酸具有结构性作用并有助于维持这些关键的残基处于正确的相对位置。所以，进化主要作用于一个蛋白质的结构，而不是它的序列，只要折叠是保守的，许多序列水平的变化就能容忍。总体的结果是在进化过程中蛋白质结构比序列更高度保守。

结构保守性的实际后果是蛋白质结构比较能够揭示遥远的进化关系，因此有助于确定不典型蛋白质的功能，甚至是当所有序列相似性的痕迹全部消失时。一个实例是 AdipoQ的特性描绘，此蛋白由脂肪细胞分泌，最初功能未知。这个蛋白的结构分析显示出与 chemokines 的肿瘤坏死因子（TNF）家族间清晰的、明确的关系，因此暗示 AdipoQ 也是一个信号蛋白。AdipoQ 与 TNFα 的结构关系见图 19.14，同时还有基于保守结构的五个超家族成员的多重序列比对（Shapiro and Scherer，1998）。

可用于蛋白质结构的两两比较算法

上面讨论的例子表明一个相对直接的注释任一孤独基因的方法就是获得编码假设蛋白的结构，并以类似于序列比较的方式与已知结构进行比较。为了达到这一目的，必须得到假设蛋白的结构（框 19.5），并应用生物信息学资源，将此结构与已经知道的蛋白

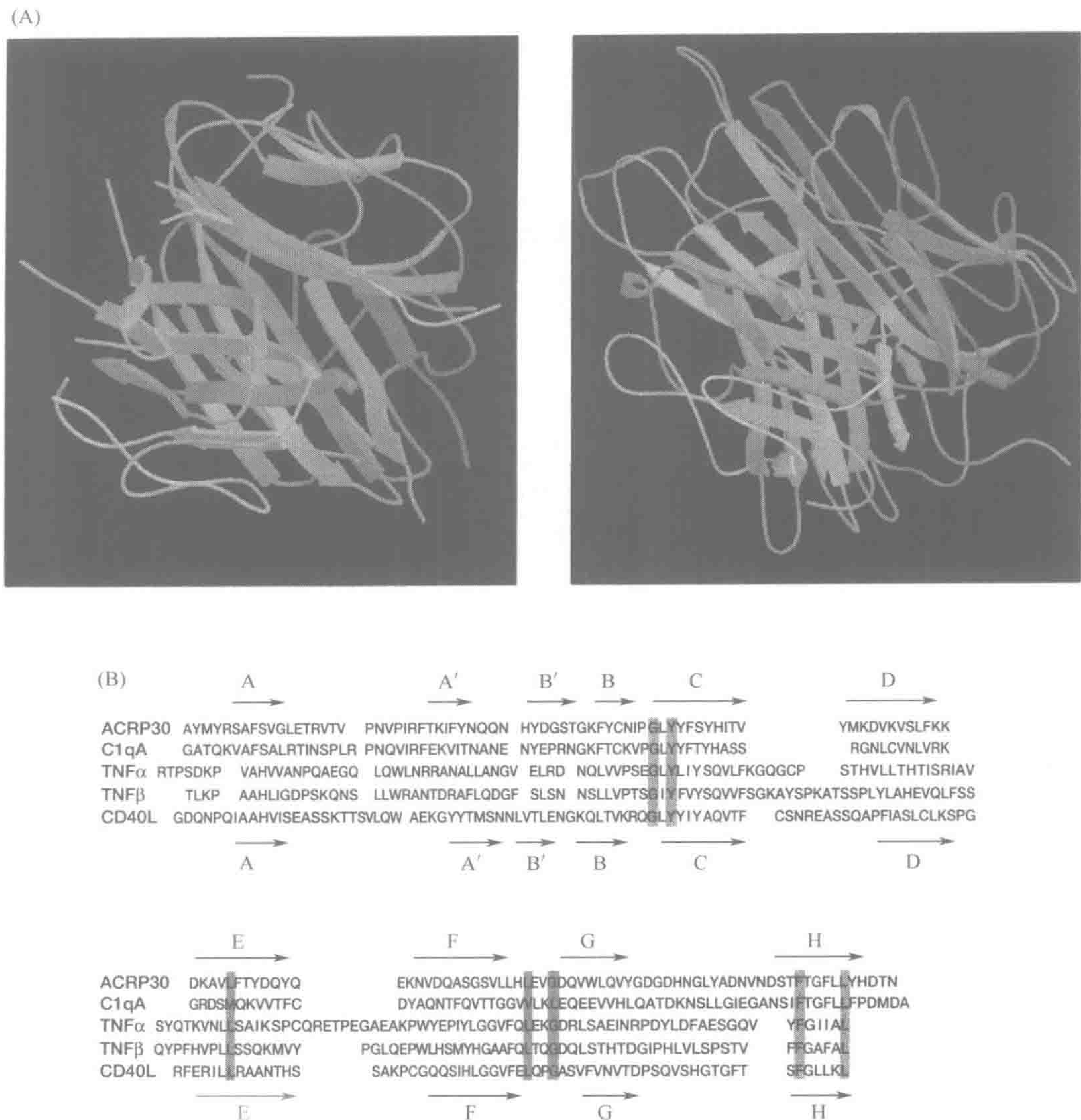


图 19.14 没有检测到序列同源处根据结构相似性的功能注释

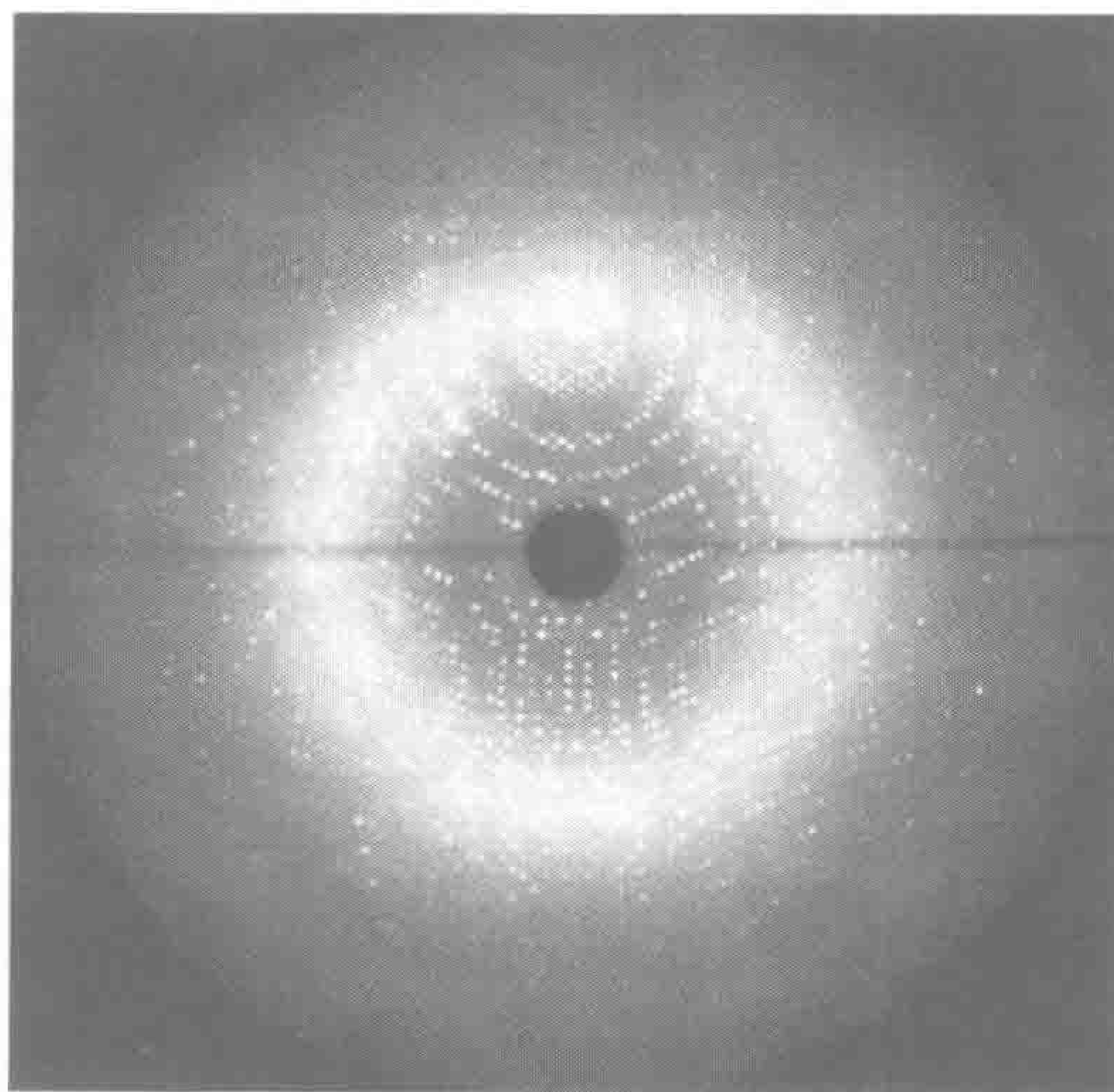
(A) AdipoQ 和 TNFα 的条带图比较。结构相似性等同于 TNF 家族内结构相似性。(B) 几个 TNF 家族成员（CD40L、TNFα 和 TNFβ）与 C1q 家族的两个成员（C1qA 和 AdipoQ）间以结构为基础的序列比对。高度保守的残基（在至少四个蛋白质中出现）以阴影部分表示，箭头所示为蛋白质 β 折叠股区域。AdipoQ 和 TNF 蛋白间几乎没有序列相似性（如 AdipoQ 和 TNFα 间 9%一致性），所以 BLAST 检索不能确定它们的关系。可是，结构比对显示保守的残基模式，这能用于在孤独基因中识别新的家族成员（节 19.2.2）。经 Elsevier Science 允许再引自 Shaprio and Schere (1998)。

质结构相比较。最大的蛋白质结构储存库是**蛋白质数据银行**（Protein Databank，PDB），该数据银行由 Rutger 大学结构生物学研究协作组维护（http://www.rxsb.org）。数据以与每个原子位置坐标的列表平面文件形式存储。这称为 **PDB 文件格式**（PDB file format）。

框 19.5　蛋白质结构的确定

通过 X 射线晶体学解释蛋白质结构

X **射线晶体学**（X-ray crystallography）利用 X 射线从蛋白质晶体散射或衍射的事实，散射特性依赖于晶体内原子组成。衍射的 X 射线能相互正向或负向干涉，在检测器上产生被称为**反射**（reflection）的形式。



蛋白质磷酸酶的 X-射线衍射图像

英国剑桥大学分子生物学 MRC 实验室 Daniela Stock 惠赠

应用一种被称为**傅里叶转换**（Fourier transform）的数学函数，能够从这些数据构建电子强度图谱。傅里叶转换中使用的数据越多，得到的结构模型越精细、越精确。精确的结构判定需要一个秩序良好、强烈地衍射 X 射线的晶体。这是结构判定一个明显的瓶颈，因为蛋白质晶体非常难于生长，即使具有自动化晶体工作站的可用性，该工作站可在轻微不同的条件下平行开展几千个反应。传统的 X 射线源可能用于结构分析，但 X 射线**同步加速器放射源**（synchrotron radiation source）更受人喜爱，因为它们能产生更高强度的 X 射线。

来自衍射数据的电子强度图谱的构建依赖于三种信息：入射 X 射线的波长（已知）、振幅和散射 X 射线的相。不幸的是，只有散射的振幅能够从反射的模式中确定，振幅必须通过进一步的**同晶型晶体**（isomorphous crystal）衍射实验来计算。同晶型晶体是相同结构的晶体，但加入更重的原子，就产生了另一种衍射模式。达到这个目的的标准方法是将晶体在一重金属盐溶液中浸泡，使得重金属原子扩散至原来由溶剂占据的空间。金属原子比通常出现在蛋白质中的轻原子衍射 X 射线更强。通过比较几种不同的同晶型晶体产生的反射光（一个称为**多重同晶型置换**，multiple isomorphous replacement 的过程），就能够计算出重原子的位置，进而使得我们能够推导出未替换晶体的衍射相。

另一种替代技术是利用**反常散射**（anomalous scattering）的现象，即当一个蛋白质晶体的金属原子受到接近于它们的吸收临界的 X 射线攻击时产生独特的衍射模式。反常散射的大小随入射 X 射线波长的变化而变化，所以一类含有金属的晶体能被几种不同波长轰击，得到不同的、能够计算散射相的衍射模式。这就是诸如**具有反常散射的单个同晶型置换**（single isomorphous replacement with anomalous scattering，SIRAS）和**多波长反常散射**（multiple wavelength anomalous scattering，MAD）技术的基础。在后一种技术中，蛋白质在细菌中表达，并加入金属替代的氨基酸硒

框 19.5　蛋白质结构的确定（续）

代蛋氨酸。在每个例子中，由于反常散射所导致的反射光强度的差异非常小，所以同步加速器放射源以及反射数据的精确记录是必需的。最后，电子强度图谱构建成一个结构模型。这需要一条更重要的信息——氨基酸序列——因为仅利用衍射数据来辨别氨基酸侧链是困难的。

应用核磁共振波谱学解释蛋白质结构

一些蛋白质不能晶体化，但是，如果它们相对较小，就可能使用**核磁共振波谱学**［nuclear magnetic resonance（NMR）spectroscopy］来解释它们的结构。直到最近，这个技术仍限定于15kDa或更小的蛋白质，但是目前数据分析的进步，使得可被研究的蛋白质最大达到100kDa大小。这一技术利用了当暴露于某一特定频率的无线电波时，某些原子核在一个应用磁场中具有在磁自旋状态间转换的倾向。当原子核弹回它们的初始方位时，它们能够发射出可以测量的无线电波。发射的无线电波的频率依赖于原子的类型和它的化学组成。例如，一个甲基基团内的氢原子核将产生与一个芳香环的氢原子核不同的信号。通过改变应用的无线电波的性质，就可能确定两个核在空间上有多么靠近在一起，即使它们不是由共价化学键连接。这称为**核欧沃豪斯效应**（nuclear Overhauser effect，NOE）。

从包括NOE信息的二维NMR光谱中，我们可能计算出哪些核是共价地结合在一起，哪些核在空间上靠近在一起。结合蛋白质氨基酸序列，就产生一个距离约束列表，首先描述蛋白质二级结构元件（这些具有非常特殊的距离关系，节1.5.5），然后发展四级结构模型。一般地，有几个四级结构的变异体同样适合于距离数据，所以它们作为一个模型整体存放，而不像在X-射线晶体学中是一个精确结构。

蛋白质结构预测

虽然解释蛋白质结构问题的技术进步显著，但它仍是一个劳动密集和昂贵的过程。尽管有点不太精确，但可选择的一种方法是应用生物信息学方法预测蛋白质结构（Jones，2000）。目前，我们可能相当准确地预测假设蛋白质的二级结构，但是四级结构的预测需要一个基于模型的模板结构。

根据发生于特殊二级结构的特定氨基酸的倾向能够预测蛋白质二级结构。某些氨基酸，如谷氨酸具有螺旋化的倾向（即它们富含于α-螺旋中）。其他氨基酸，像缬氨酸，具有成股的倾向（即它们富含于β折叠股和β片层中）。甘氨酸和脯氨酸不常见，因为根本就很少在二级结构中见到它们。实际上，常常在螺旋和折叠股的末端看到它们，似乎具有"结构终止子"的作用。具有螺旋和折叠股倾向的一串残基的出现强烈提示在蛋白质中这样一个结构的存在。然而，建立在单个蛋白质基础上的二级结构预测并不可靠，因为存在所有的氨基酸都出现在所有类型二级结构中的个别例子。多重比对能通过鉴定有利于螺旋或折叠股形成的保守性残基模块来去除这个不确定性。最复杂的算法，如PSI-PRED（Jones，2000）采用多重比对，也加入了来自数据库的进化和结构信息，增加了它们预测的准确性。

四级结构比二级结构的预测要难得多，因为有上千万种不同的折叠任何既定的线性氨基酸链的方法。如果为了使其成为现实而在模型中加入足够的溶剂分子，但没有已知蛋白的行为知识，系统就变得太复杂而不能研究。因为这个原因，从头开始的预测方法没有实际使用。然而，如果能够得到一个关系密切的蛋白质的结构，它就能用作模板，构建查询序列的一种结构模型。这称为**比较建模**（comparative modeling）或**同源建模**（homology modeling），通常在两个序列显示大于25%一致性时起作用。更遥远的关系能够通过串线方法建模，在此方法中，假设蛋白的序列与蛋白质数据银行的序列相比较，试图找到一致的折叠。可能常常发现折叠与蛋白质核心匹配，而蛋白质核心高度保守但外部环袢变异极大。采用所谓的**散件算法**（spare parts algorithm），可能在其他蛋白质上找到相似的环袢。因此，结构模型作为一系列碎片从与已知结构一致的序列中建立。串线以及比较建模的最后步骤都是模型的精细化，这涉及移动侧链以避免冲突，并使结构的总自由能最小化。

几个计算机程序可免费从互联网获得，该程序将 PDB 文件转换为三维模型（如，Rasmol，MolSript，Chime）。此外，大量的算法已经编写用于蛋白质结构比较（Sillitoe and Orengo，2002）。一般来说，这些依据两个原理之一进行工作，尽管一些更新的程序同时应用了两个元素：

- **分子间比较**（intermolecular comparison）：将两个蛋白质的结构叠加，算法试图使叠加原子之间的距离最小化（图 19.15A）。用于测量结构间相似性的函数通常是**均方根差**（root mean square deviation，RMSD），即当量原子间平方距离均值的平方根（图 19.15A）。当蛋白质结构变得更加相似时，RMSD 下降；如果两个相同结构叠加，则 RMSD 是零。这种算法的例子包括 Comp-3D 和 ProSup；
- **分子内比较**（intramolecular comparison）：两个蛋白质的结构并排比较，算法测量的是每个结构中当量原子间的内部距离，并鉴定与这些内部距离最密切匹配的比对（图 19.15B）。这种算法的一个例子是 DALI。采用两种方法的算法包括 COMPARER和 VAST。

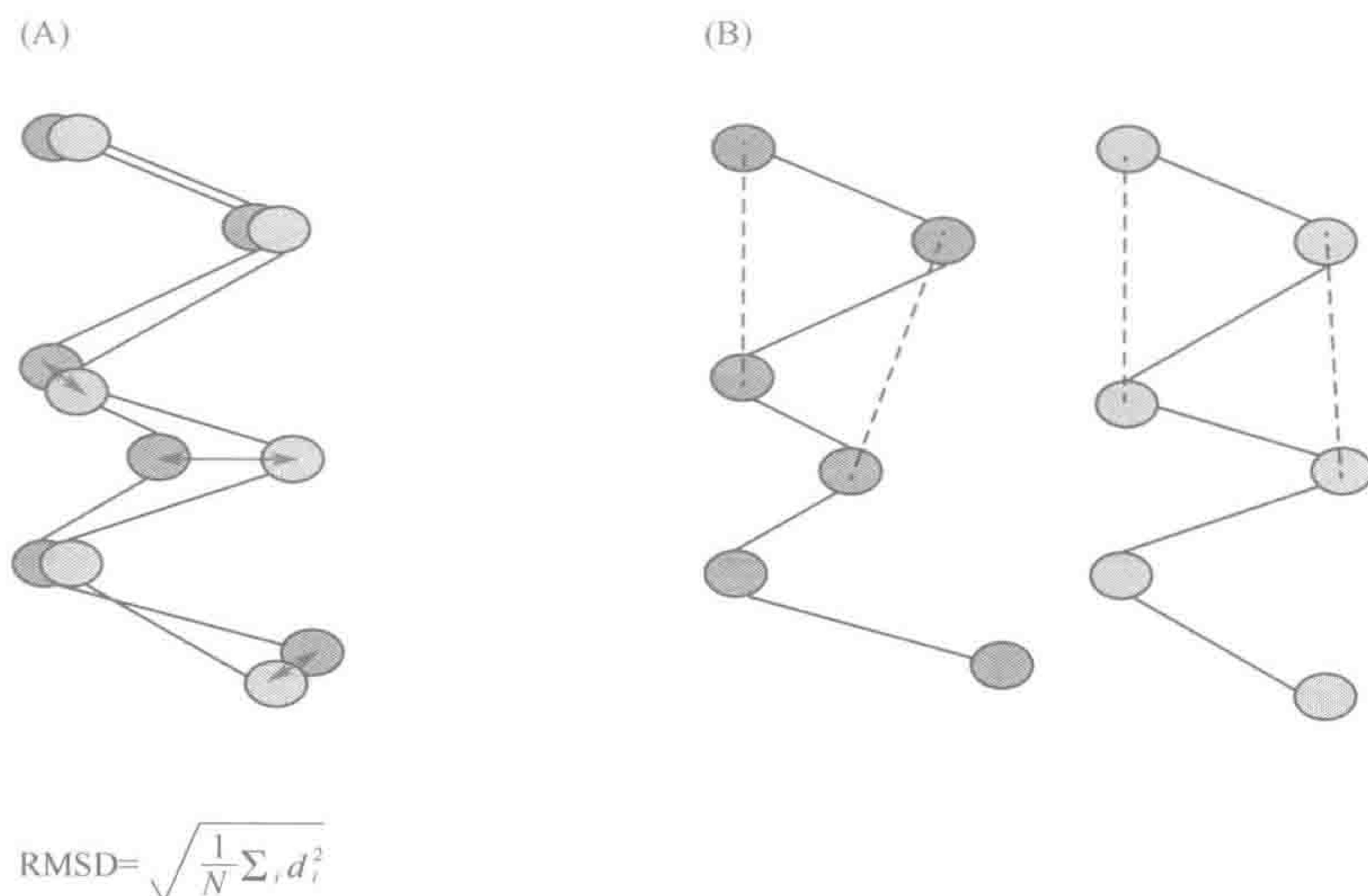


图 19.15 蛋白质结构比较，圆圈代表每个氨基酸残基的 Cα 原子，直线代表多肽骨架的空间路径

（A）分子间比较涉及蛋白质结构的叠加，以及计算叠加结构中当量原子间距离的计算（双向箭头所示）。这些距离用于计算均方根差（RMSD），如公式所示（R 是 RMSD，d 是第 i 对叠加的 Cα 原子间距离，N 是排列原子的总数）。在许多残基基础上计算得到的一个 RMSD 数值是明显保守四级结构的证据。（B）分子间比较涉及并列分析，基于每个结构中当量原子之间的比较距离（虚线所示）。

结构基因组学（结构蛋白质组学）目的是解释覆盖折叠空间的一套代表性蛋白质的结构

人类基因组有大约 30000 个基因，但它们中多数能被归类为具有相似序列的家族（种间同源基因）。像上面讨论的，相似的序列暗示了相似的结构，所以每个基因家族中

的不同成员可能编码具有相同折叠的蛋白质。甚至没有显示出明显的序列同源性的基因家族能够编码结构相似的蛋白质，如上面 AdipoQ 和 TNF 家族所显示的。考虑到多个结构域蛋白质的复杂性，据估计现存的蛋白质折叠较少，只有 1000 种。这意味着如果每个折叠家族中的一个代表性蛋白质能从结构上解析，那么就有可能利用结构数据注释所有的孤独基因。

世界各地已经开始实行几个小规模试验程序，以在总体水平上解释蛋白质结构，尝试覆盖“折叠空间”，并获得一套代表性蛋白质的结构（Brenner，2001；Norin and Sundstrom，2002；Zhang and Kim，2003）。在大多数计划中，采用多波长反常散射（MAD）技术（框 19.5 和 Heinemann *et al.*，2001）进行高通量的 X-射线晶体学分析。在这些程序中，一个常见的主题是漏斗效应，即在分析的每一个阶段都有一定比例的蛋白质流失。对于每 25 个表达的蛋白质，只有五个能产生晶体，但仅有一个能产生有用的衍射数据。由于某些程序专注于与疾病相关的人类或病原体蛋白质，所以大多数涉及来自具有小型基因组的细菌（如流感嗜血杆菌，*Haemophilus influenzae*）或嗜热细菌（如詹氏产甲烷球菌，*Methanococcus jannaschii*、嗜热产甲烷杆菌，*Methanobacterium thermoautotrophicum*）的蛋白质。这是因为具有小型基因组的细菌也具有小型蛋白质组，但仍然包含所有蛋白质家族的代表，而来自嗜热细菌的蛋白质在大肠杆菌中表达时应该更稳定。这些小规模试验程序的输出结果提示大约 70%的假设蛋白质含有已知的折叠，理论上可以依据结构数据进行注释，而 30%含有功能仍未知的全新折叠。复杂性包括难以提供蛋白质结构和“俄罗斯玩偶效应”的严格定义，后者在确定无误的折叠类型中有连续范围的中间结构。这些问题在框 19.6 中讨论。

框 19.6　蛋白质的结构分类

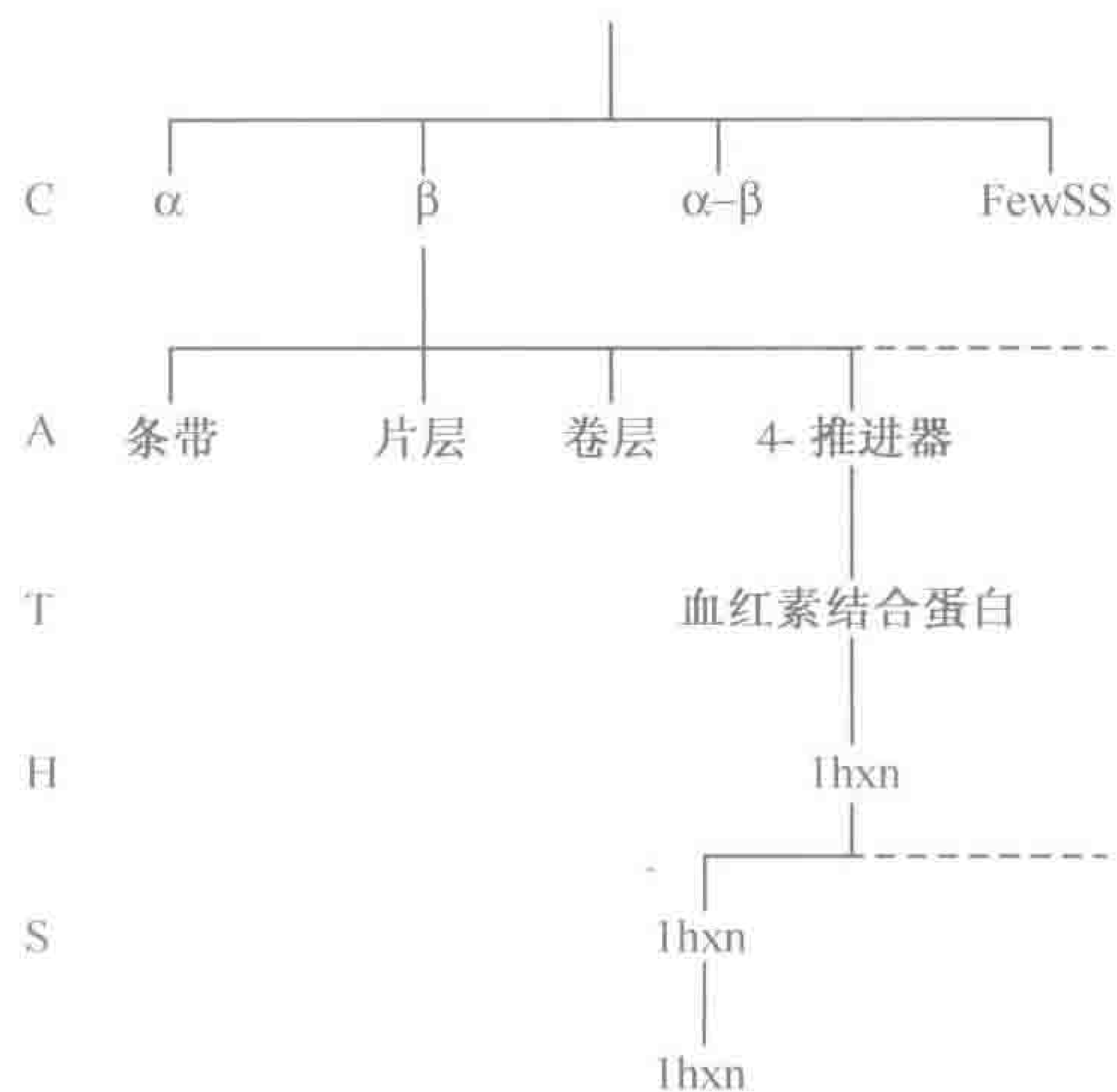
基于蛋白质结构的功能性注释需要一个严格的、标准化的系统用于不同结构的分类。几种不同的等级分类系统是可用的，在这些系统中，蛋白质依据其含有的各种二级结构的比例首先划分为大体类别，然后依据那些结构如何排列，连续地进行更专业化的分组（Pearl and Orengo，2002）。这些系统在数据库中贯彻执行，例如，SCOP（蛋白质结构分类，Structural Classification of Proteins）CATH（分类、结构、拓扑和同源性超家族；Class，Architecture，Topology，Homologpus superfamily）以及 FSSP（基于蛋白质结构－结构比对的折叠分类，Fold classification based on Structure-Structure alignment of Proteins）。

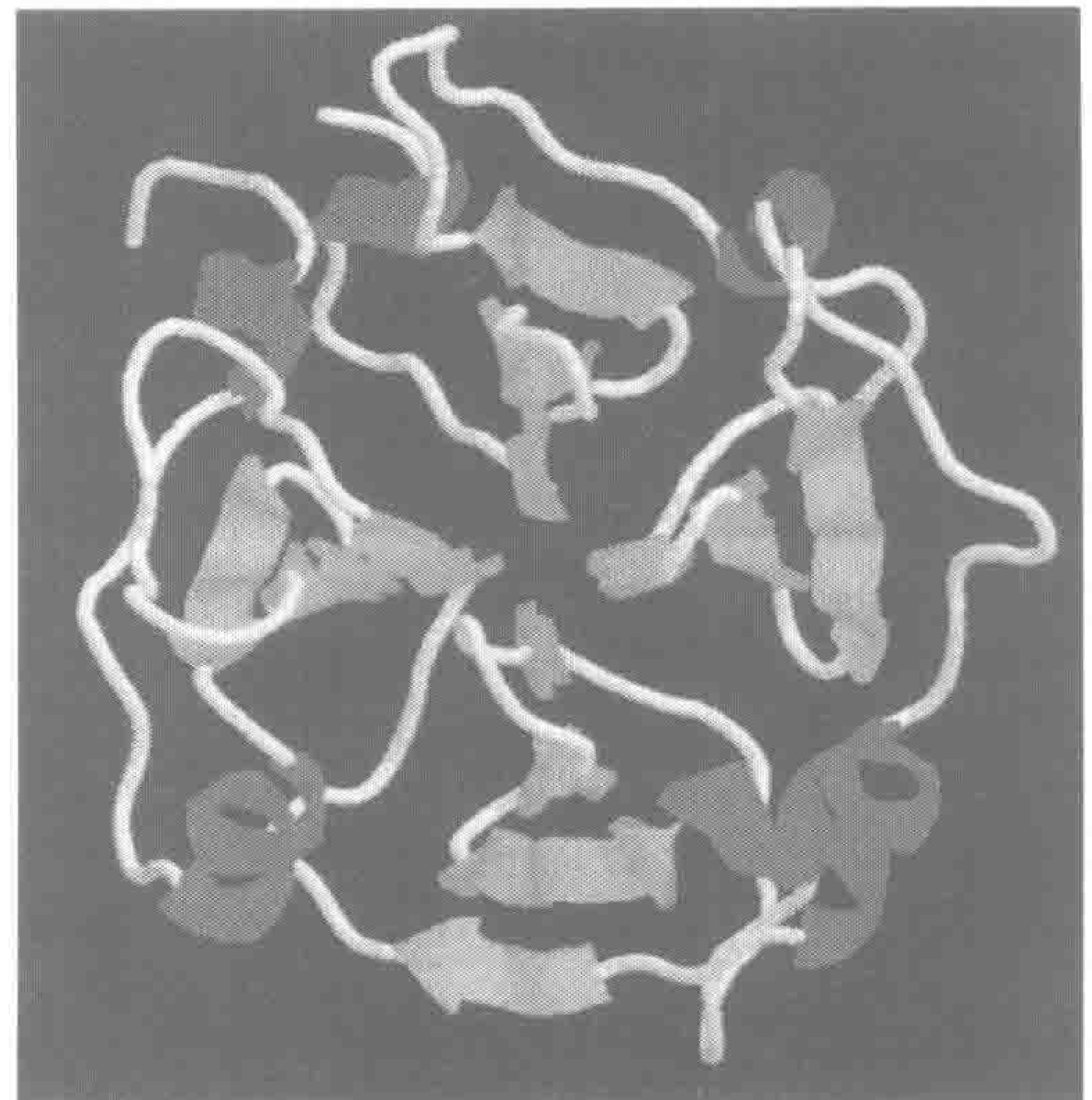
这些数据库进行分类的方法不同。例如，FSSP 系统是应用 DALI 程序进行完全自动化的结构比较来实施分类。CATH 是半自动化的，采用手工排列的自动比较。SCOP 是一个手工分类系统，建立在进化关系以及几何学标准基础之上。毋须吃惊，当应用另一种系统时，同一蛋白质可能分类不同（Hadly and Jones，1999）。在等级的较高水平，具有广泛的普遍一致性，但是当追求更细致的分类时，总会遇到问题，因为这依赖于不同分类系统中用于识别折叠基团的阈值。

导致蛋白质结构分类混淆的其他问题包括：

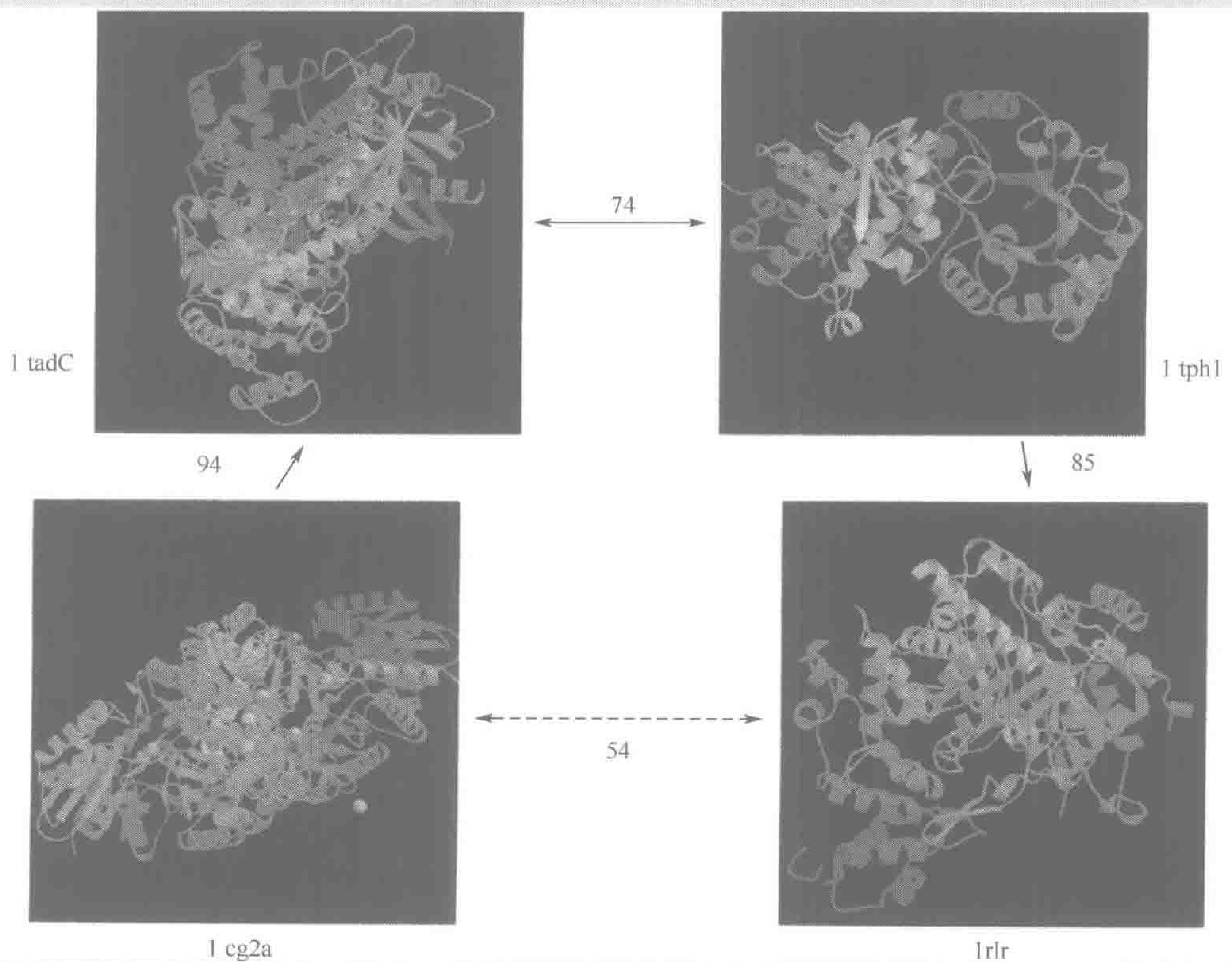
- 所谓的**超折叠**（superfold）的存在：在许多具有多种多样四级结构的蛋白质中发现。必须区分同源性结构（来源于一个共同的进化祖先）和类似结构（独立进化，但已经集中）；
- 同一蛋白质家族不同成员间折叠结构的变异，能够导致识别同源性关系的失败；
- **俄罗斯玩偶效应**（Russian doll effect），它描述了折叠基团间一个结构的连续统一体，所以将一种结构分配到一个特定的类别就变得非常主观。

框 19.6 蛋白质的结构分类（续）





CATH 数据库中等级性结构分类的一个实例



俄罗斯玩偶效应（Russian doll effect）

显示了在空间折叠上表现为连续性结构变化的四个蛋白质。每个蛋白质与其最近的邻居共享至少 74 个结构上相同的残基，但当直接比较两个极端蛋白时，它们显示仅有 54 个结构上相同的残基。图解：1cg2A，羧肽酶 G2；1tadC，转导蛋白-K；1thp1，磷酸丙糖异构酶；1rlr，核糖核酸还原酶蛋白 R1。依据 Domingues *et al*. 2000，FEBS Letters，476，98～102，Figure 2.

19.4.5　研究单个蛋白质相互作用的多种不同方法

甚至在经过蛋白质的序列、结构和表达谱的综合性研究之后，蛋白质功能可能并不清楚。在这些情况下，鉴定那些与其特异性相互作用的蛋白质可能更有益处，特别是如果这些证明是已被研究透彻、功能已知的蛋白质。例如，如果一个未定性的新蛋白，显示出与 RNA 剪接所需的其他蛋白质相互作用，那么这个新蛋白可能在同一过程中发挥作用。通过这种方法，蛋白质能够与细胞中的功能网络相联系。

大量方法可用于单个蛋白质相互作用的研究，包括遗传的、生化的和物理技术（Phizicky and Fields，1995）。遗传学方法仅能应用于模式生物，诸如果蝇和酵母，而生化和物理方法能够直接用于人类细胞。一个经典的生化方法是**亲和层析**（affinity chromatography），在这种方法中，一个特定的诱饵蛋白固定于一层析柱的支持介质上，再加入细胞裂解物。与诱饵相互作用的蛋白质就保留在柱中，而其他蛋白质被洗掉。通过提高缓冲液的盐浓度洗脱相互作用的蛋白质。另一个生化方法是**免疫共沉淀**（co-immunoprecipitation），该方法是将某一特定诱饵蛋白的特异性抗体直接加入到细胞裂解物中，结果导致抗体—诱饵蛋白复合体沉淀。任何能与诱饵相互作用的蛋白质均被免疫共沉淀。亲和层析和免疫共沉淀都能够用于分离完整的**蛋白质复合体**（protein complex），使得不同的组分通过 MS 来鉴定（节 19.4.2）。这是相互作用蛋白质组学中一个主要技术平台。它已经广泛地用于蛋白质复合体的定性，如核糖体、后期启动复合体，核孔复合体和信号复合体，近期已在基因组的规模使用（例如，Gavin *et al.*，2002；Ho *et al.*，2002；图 19.16）。每个复合体内部特异性相互作用能够通过化学交联进行研究。

19.4.6　应用基于文库的方法进行高通量相互作用筛查

如果高通量相互作用筛查方法提供蛋白质和编码它们的基因之间一个直接的连接，那就会获益匪浅。这由于以文库为基础的相互作用筛查方法的发展而被提出。原则上，能够利用一个蛋白质诱饵代替一个 DNA 探针筛查任何标准的 cDNA 表达文库。然而，实际上这是艰巨的，因为同一个文库必须进行很多次筛查来描述不同诱饵的相互作用。两种新技术已用于蛋白质相互作用的筛查，这将在下面讨论。

通过噬菌体展示进行的相互作用筛查涉及筛选固定于微滴定皿孔中的诱饵蛋白与表达于重组噬菌体表面的相互作用蛋白质之间的相互作用

噬菌体展示是一种表达克隆形式，在此方法中，外源 DNA 片段插入到噬菌体外壳蛋白基因中（节 5.6.2；Burton，1995）。随后，重组基因能够作为融合蛋白表达，整合至病毒粒子并在噬菌体表面展示（图 19.17）。融合噬菌体将会与任何与其表面展示的外源成分相互作用的蛋白质结合。通过创建一个噬菌体展示文库使蛋白质组中每种蛋白质都展示在噬菌体表面，就可以进行相互作用筛查。随后，微滴定板的孔中包被特定的、感兴趣的诱饵蛋白，并将噬菌体展示文库导入每个孔中。表面带有相互作用蛋白质的噬菌体将继续保持结合在孔的表面，而那些没有相互作用的蛋白质被洗脱。此技术的一个特殊的优点是保留的噬菌体展示相互作用蛋白能够从孔中洗脱，并用于感染大肠杆

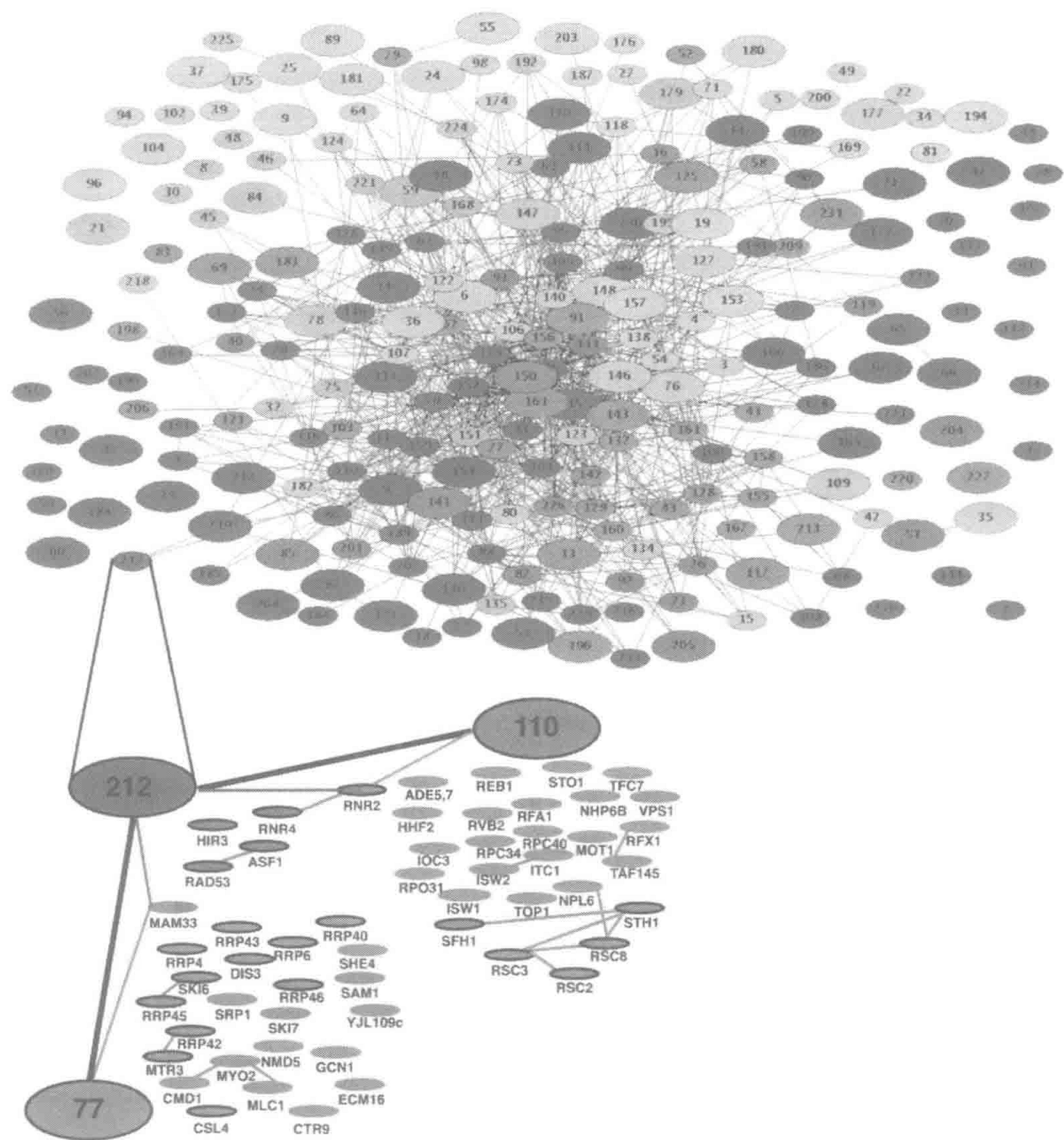


图 19.16　在酵母中采用质谱方法系统性分析蛋白质复合体发现的蛋白质相互作用的网络

一个代表酵母蛋白质复杂网络的图解，是通过连接共享至少一个蛋白质的复合体而建立的（为保证清晰性，共享超过九个蛋白质的复合体已被忽略）。在上面的样板中，单个复合体的细胞作用标记彩色如下：红色，细胞周期；深绿，信号传导；深蓝，转录、DNA 维持和染色质结构；粉红，蛋白质和 RNA 转运；橘黄，RNA 代谢；浅绿，蛋白质合成和更新；棕色，细胞极性和结构；紫色，中间产物和能量代谢；浅蓝，膜生物发生和运输。下面的样板是一个复合体（TAP-C212）通过共享组分（红线表示已知的物理相互作用）与其他两种复合体（TAP-C77 和 TAP-C110）相连接的例子。经 Nature Publishing Group 和 Cellzome AG 允许，在 Gavin 等.(2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415，141～147 之后。

菌，结果导致相应 cDNA 序列的大量扩增，随后能够获得 cDNA 序列并通过数据库检索用于鉴定相互作用蛋白。噬菌体展示的缺点包括体外实验的形式，以及只有短肽序列

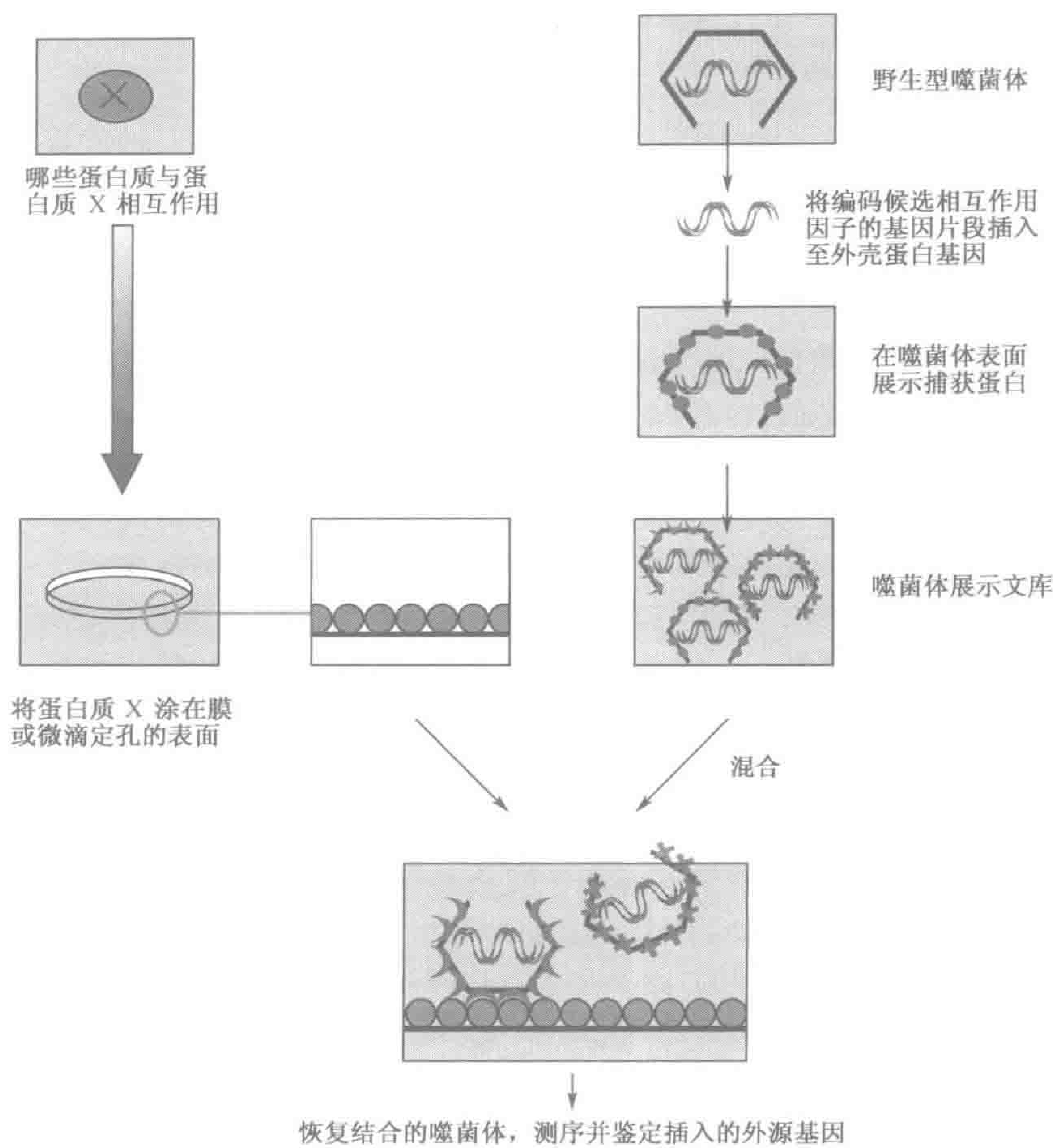


图 19.17　应用于高通量相互作用筛查时噬菌体展示的原理（一般原理见图 5.22）
相互作用因子寻找的诱饵蛋白能够为固定在微滴定孔或膜的表面。然后蛋白质组中所有其他的蛋白质通过克隆导入至噬菌体外壳基因，表达于噬菌体的表面，产生噬菌体展示文库。随后，孔或膜被灌注噬菌体文库。对于任何既定的诱饵蛋白（X），携带相互作用蛋白质的噬菌体将被保留，而那些展示没有相互作用蛋白质的噬菌体将被洗脱。结合的噬菌体能够在高盐缓冲液洗脱，并用于感染大肠杆菌，产生大量含有相互作用蛋白质的 DNA 序列的噬菌体颗粒。

能够展示在噬菌体表面而不破坏其复制周期的事实。该技术的这两个特点可能阻碍特定相互作用的发生。

除了相互作用筛查之外，噬菌体展示方法也有益于其他几方面的应用，包括：

- 抗体工程：噬菌体展示证明是一种有力的抗体替代性资源（包括人类抗体）。它能够绕过免疫实验，甚至是杂交瘤技术（Winter *et al*.，1994；节 21.3.4）；
- 常用蛋白质工程：作为从一个突变体文库中筛选所需变异体的一种途径，噬菌体展示是随机诱变程序一个有力的附属物。

酵母双杂交系统具有任何相互作用筛查技术的最高通量，涉及来自独立成分的一个功能性转录因子的组装

在相互作用筛查中，使用最广泛的文库方法是酵母双杂交系统，也称为相互作用捕获系统（Fields and Sternglanz，1994）。这个方法既可以鉴定与待研究蛋白结合的蛋白质，也能够用于描述相互作用中关键的结构域和残基。确实相互作用的蛋白质通过它们组装一个有活性的转录因子进而激活一个报道基因和/或选择性标记来检测。酵母双杂交方法的关键是对转录因子由两个独立结构域组成的观察：一个是DNA结合结构域，一个是反式激活结构域。在大多数天然的转录因子中，DNA结合和激活的结构域是同一多肽的一部分（节10.2.4）。然而，一个有活性的转录因子也能够由两个相互作用的、携带独立结构域的蛋白质组装。双杂交系统及其衍生技术的目的是将一个靶蛋白作为诱饵，用于一个相互作用蛋白质的特异性识别，而这个蛋白质与一个必需的转录因子成分融合。

为了应用双杂交系统，采用标准的重组DNA方法产生一个编码待研究**诱饵蛋白**（bait protein）的融合基因，并与一个转录因子的DNA结合结构域相偶联（图19.18）。这个基因转化的细胞与融合基因文库转化的细胞相接合，融合基因文库中的cDNA序列与反式激活结构域的编码序列（**捕获构建体**，prey construct）偶联。靶细胞也被改造，带有能被组装转录因子激活的报道基因和/或选择性标记基因。在二倍体酵母细胞中进行相互作用检测。如果诱饵和捕获不能相互作用，则两个转录因子结构域保持分离，标记基因失活。可是，如果诱饵和捕获的确相互作用，则转录因子组装，随后标记基因被激活，产生可视性标识和/或含有相互作用蛋白质的酵母细胞的选择性繁殖。从这些细胞中能够鉴定捕获构建体的cDNA序列。

为了进行单个蛋白质（单个诱饵）的分析，开发了酵母双杂交筛查技术，但是在最近几年中，这个技术已被更大规模地应用，达到高潮的是酵母中全部40 000 000种可能的蛋白质相互作用的详尽研究（Uetz *et al.*，2000；Ito *et al.*，2000，2001）。该技术有两个变异体，一个是系统地生产和杂交明确的诱饵和捕获对象（**矩阵方法**，matrix method）；另一个是由随机的cDNA片段代表捕获或者诱饵与捕获（**随机文库方法**，random library method）（图19.19）。矩阵方法是详尽的、系统的，但是没有随机文库方法可信，因为蛋白质组中的每个蛋白质由一单个的构建体代表，而这些构建体一定是通过PCR独立地生成。文库方法涉及更少的工作量，更有优势，因为每个捕获是由多个重叠的克隆来代表。这就降低了假阴性的可能性，因为能够使用同一捕获构建体的变异体；同时也降低了假阳性的水平，因为与代表同一捕获的不同构建体的独立匹配是由于能够检出任何相互作用而提高了可信的水平。但是总的来说，大规模筛查的结果必须要谨慎解释，因为有多种因素能够导致假阳性和阴性结果的出现（综述见Legrain *et al.*，2001）。虽然大多数整体双杂交研究集中于微生物，但最近的一个小规模试验程序显示该技术可以同样地应用于哺乳动物蛋白质组（Suzuki *et al.*，2001）。

除了基本的双杂交系统之外，还有衍生的具有更专业化应用的方法。这些包括下面的方法：

- **单杂交系统**（one-hybrid system），用于检测蛋白质与特定DNA或RNA序列的相互作用；

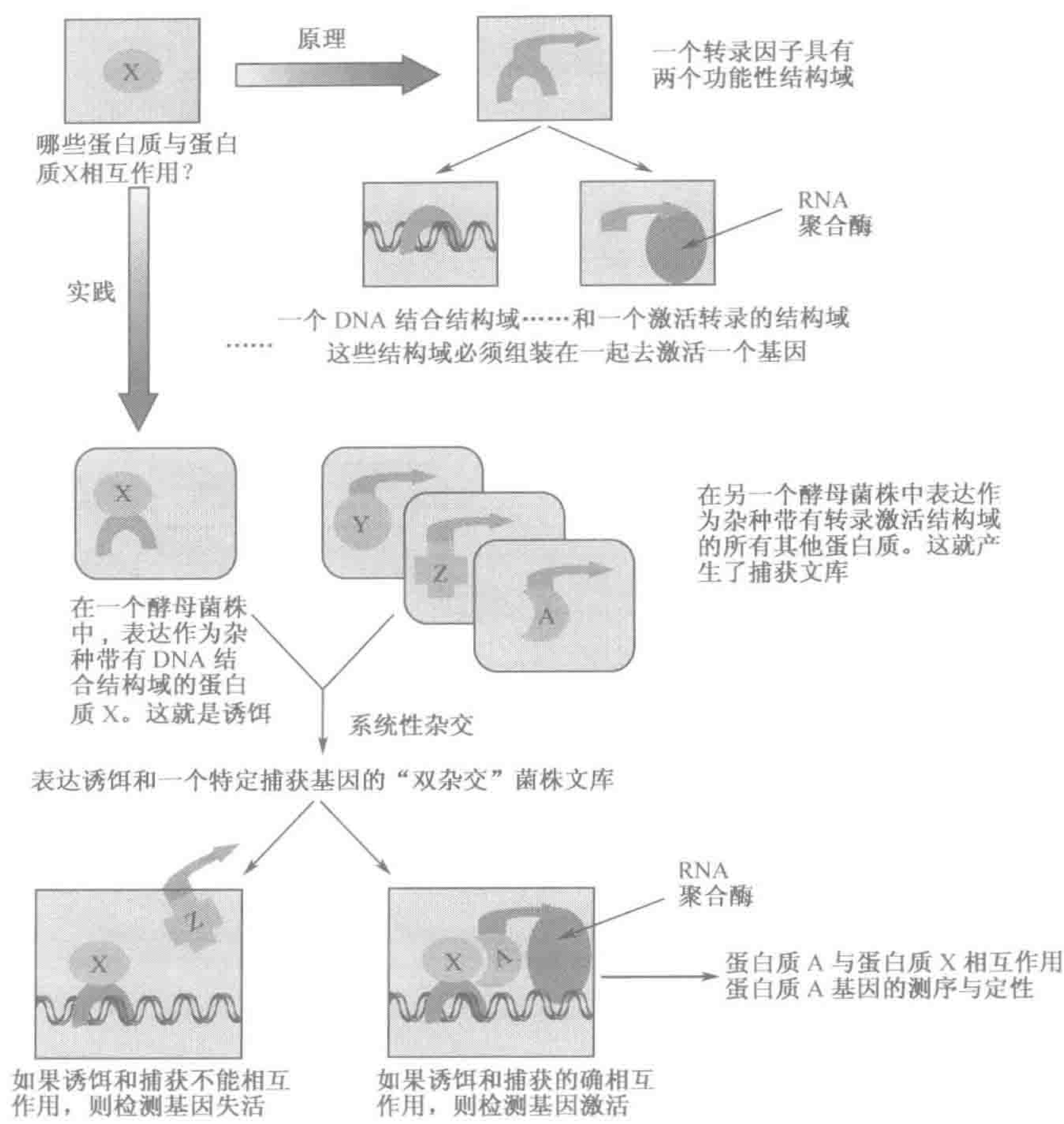


图 19.18　酵母双杂交系统的原理和实践

通常转录因子由两个功能上独立的结构域组成，一个用于DNA结合，一个用于转录激活。它们并非共价地结合在一起，但是能够被组装成一个二聚体蛋白。可利用这一原理鉴定蛋白质相互作用，诱饵蛋白与DNA结合结构域融合，表达于一种酵母菌株；候选的捕获基因表达于另一种菌株，并与一个反向激活结构域融合。当两个菌株接合时，仅在诱饵和捕获相互作用时，才能组装功能性转录因子。这可通过包括一个由杂种转录因子激活的报道基因来检测。尽管原理简单，但该技术易于出现可靠性和可重复性问题。假阳性的产生是由于自发性自主激活（诱饵或捕获能够自主激活报道基因），黏性诱饵和捕获（与许多其他蛋白质非特异性相互作用的蛋白质）以及无关的相关作用（在正常条件下决不会彼此面对的偶然的蛋白质间的相互作用，例如那些通常发现于独立分隔空间的蛋白质）。假阴性的产生可能由于构建中的PCR错误，以及非生理性条件（实验在细胞核中发生，所以来自其他分隔空间的蛋白质无法正确地折叠和组装）。

- **三杂交系统**（three-hybrid system），用于研究更复杂的蛋白质相互作用，包括涉及RNA和蛋白质（诱饵和鱼钩）的相互作用；
- **反向双杂交系统**（reverse two-hybrid system），用一个能够被正确组装的转录因子激活的自杀基因酵母检测蛋白质相互作用的破坏。
- **分裂泛素系统**（split ubiquitin system），在此系统中，蛋白质相互作用将泛素蛋白质的两个二等分组装在一起。相互作用可通过检验诱饵蛋白被泛素降解来监测，或者

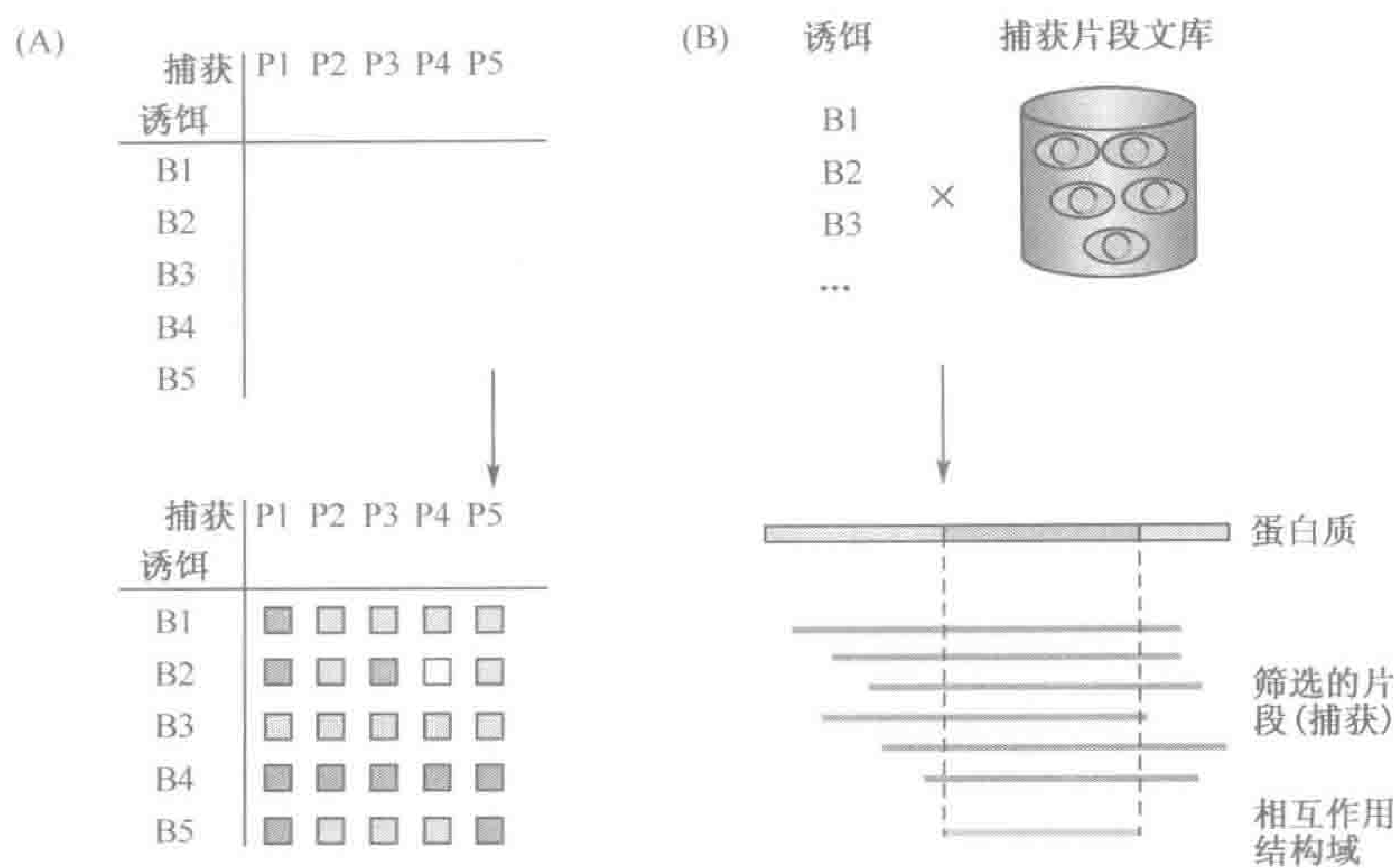


图 19.19　用于构建大规模蛋白质相互作用图谱的矩阵和文库筛查方法

(A) **矩阵方法** (the matrix approach)：该方法使用了相同的蛋白质集合（1～5）作为诱饵（B1～B5）和捕获（P1～P5）。结果可绘制成一个矩阵。自主激活因子（例如 B4）和“黏性”捕获蛋白（例如，P1 与许多诱饵相互作用）被鉴定并弃掉。最终的结果总结成一个相互作用的列表，相互作用可为异源二聚体（如 B2－P3）或者同源二聚体（B5－P5）。(B) **文库筛查方法** (library screening approach)：这个方法鉴定与一既定诱饵相互作用的每个捕获蛋白质的相互作用结构域。不考虑诱饵蛋白，黏性捕获蛋白作为常被选择的蛋白质片段而被鉴定出来。经 Elsevier Science 允许再引自 Legrain 等（2001）。

通过一个功能性蛋白质如一个转录因子的释放来监测；

- **SOS 复原系统** (SOS recruitment system)，在此系统中，将相互作用的蛋白质复原到膜上，并完成一个必需的信号通路。诱饵蛋白局限于缺少 CDC25 活性的酵母细胞的细胞质膜上。酵母 cdc 25 突变体不能存活，但只要蛋白质定位于膜上，就能够通过人类的种间同源 SOS 来修复它们。通过表达一个像 SOS 杂种的捕获文库，就能够进行检测。这个系统有益于研究转录因子的相互作用，这些转录因子在常规的双杂交系统中能够自主激活报道基因；
- 哺乳动物双杂交系统 (mammalian two-hybrid system)，能够检测具有真正翻译后修饰的蛋白质间的相互作用。

19.4.7　相互作用蛋白质组学的挑战是组装一个细胞功能性相互作用图谱

蛋白质相互作用数据提供了有益的功能性信息，而且在全蛋白质组水平，潜在地使得细胞中所有蛋白质与功能性通路和网络相联系。这一数据的消化和描述很重要，而且考虑到这个目的已经建立了几个数据库（表 19.2）。这些数据绝大多数来源于大规模 MS 和双杂交筛查，但有潜在地极大数量的数据可能涉及“隐藏”在一些可以追溯到许多年以前的科学文献中个别的蛋白质相互作用（见图 19.16 的蓝线）。提取这些信息并将其与从最近的高通量实验中获得的信息整合将是一个挑战。有趣的是，已经开发了几个生物信息学工具来查阅文献和识别指示蛋白质相互作用的关键词，以便使相互作用数据库的人类管理者能够细查这样的参考文献（Xenarios and Eisenberg，2001）。

表 19.2　存储蛋白质相互作用信息的数据库精选

数据库和注释	URL
生物分子相互作用网络数据库 (Biomolecular Interation Network Database,BIND) 列举了单个蛋白质-蛋白质相互作用、通路和复合体，还列举了蛋白质与其他分子的相互作用	http://www.bind.ca
相互作用蛋白质数据库 (Database of Interacting Proteins,DIP) 列举几千个蛋白质-蛋白质相互作用	http://dip.doe-mbi.ucla.edu
基因和基因组的京都百科全书 (Kyoto Encyclopedia of Genes and Genomes,KEGG) 关于代谢和信号通路的广泛资源，也有一节介绍蛋白质复合体结构	http://www.genome.ad.jp/kegg/

第二个问题是蛋白质相互作用网络并非描述那样简单，所以展示清晰的数据是一个很大的挑战（图 19.16）。然而，就像所期待的那样，具有相似普遍功能（如膜转运体、DNA 修复、氨基酸代谢）的蛋白质倾向于彼此发生相互作用，而不是与其他功能不相关的蛋白质相互作用。因此，通过将那些功能上相关的蛋白质分组至代表基本细胞过程的集中点，复杂图谱就能够被简化，如图 19.20 所示，值得注意的是，这些过程本身以各种方式相联系，如调控染色质结构的蛋白质与那些涉及 DNA 修复和重组的蛋白质间存在更多相互作用，而不是与那些涉及氨基酸代谢和蛋白质降解的蛋白质。这种描述方法使得蛋白质相互作用按等级方式呈现，也可以提供基准点，保证新的相互作用判定近乎合理。全部蛋白质相互作用的四分之三发生在相同功能性蛋白质组内，而其余大多数发生于相关的功能性蛋白质组内。参与不相关过程的蛋白质之间一个预料之外的相互作用被认为是可疑的，可以通过严格的遗传、生化和物理方法进行验证。

19.4.8　蛋白质与小配体相互作用的信息能够增加我们对生物分子过程的了解，并为药物设计提供一个合理的依据

如同蛋白质-蛋白质相互作用（见上文）以及蛋白质-核酸相互作用，蛋白质还与大量小分子相互作用，这些小分子作为配体、底物、运输工具（就转运体蛋白来说）、辅助因子以及变构调节因子起作用。在所有的实例中，这些相互作用的发生是因为蛋白质表面与相互作用分子在形状和化学特性上互补，这意味着结合造成自由能的降低。药物设计的根据是鉴定在细胞中与靶蛋白发生特异性相互作用，并改变其活性的分子。大多数药物在蛋白质水平发挥作用，通过与蛋白质相互作用，并以生理上受益的方式改变它的活性来实现。药物相关的副作用通常反映与其他蛋白质非特异性相互作用所造成的有害效应。我们对蛋白质结构和蛋白质与小分子相互作用了解得越多，我们就能够利用越多的信息设计更有效、更特异的药物。

蛋白质-配体相互作用能够通过高通量筛查大量化学复合物来实现，但工作量可以通过最初尝试模拟这样的相互作用以便确定适宜的**前导复合物**（lead compound）来降

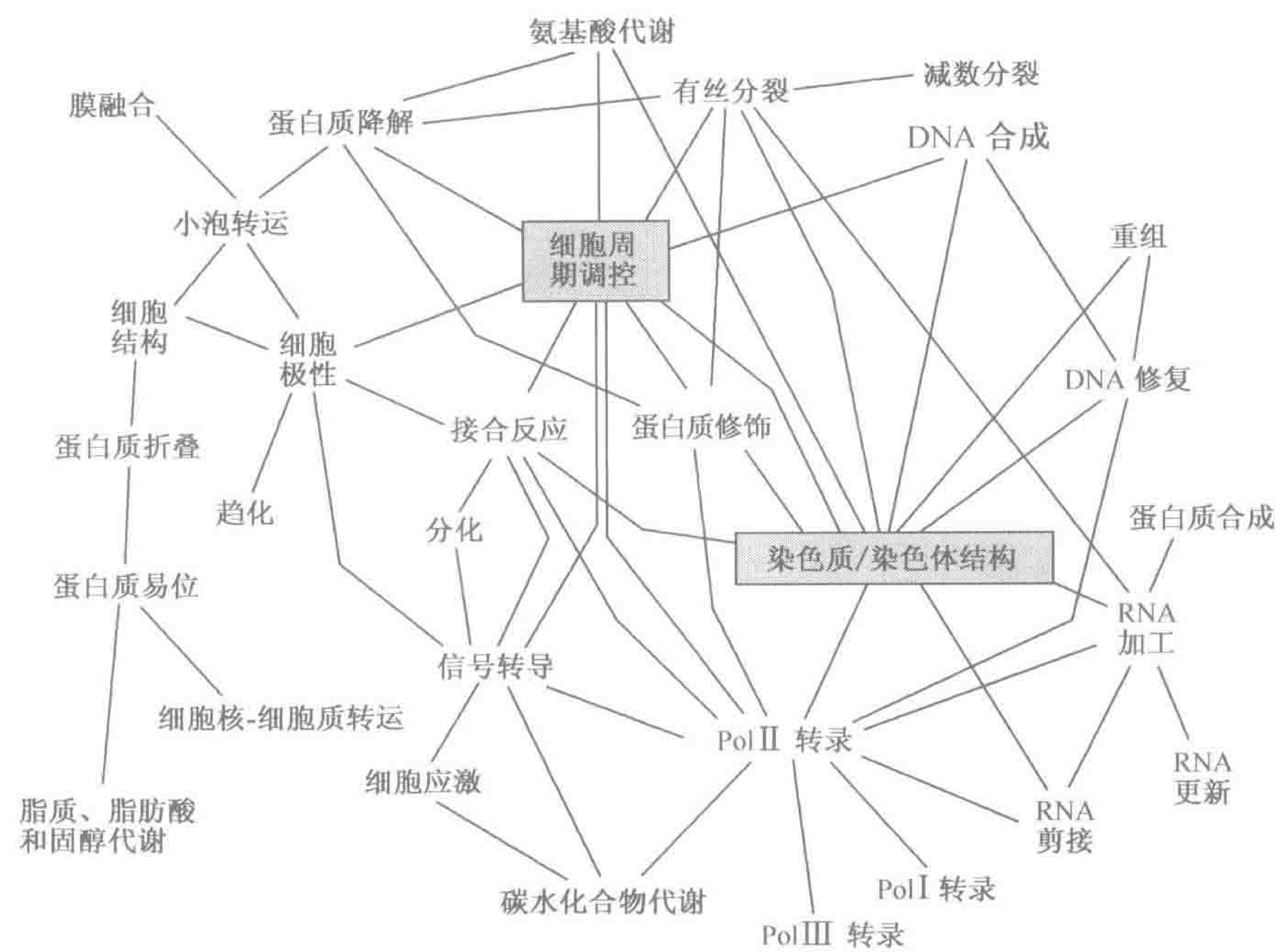


图 19.20　酵母蛋白质组简化的“功能组相互作用图”

这些数据来源于一个更为复杂的个体相互作用的图谱，大多数通过双杂交筛查获得。每条线表明在有联系的组的蛋白质之间存在 15 种或更多的相互作用。少于 15 种相互作用的关系没有显示，因为少数相互作用只发生于几乎所有组之间，通常倾向于是可疑的，即基于假阳性的结果。只包括完全注释的酵母蛋白质，而许多蛋白质是已知的，属于几个功能类型。经允许 Elsevier Science 再引自 Tucker 等（2002）。

低。如果能得到靶蛋白的结构（见上文），称为**对接算法**（docking algorithms）的计算机程序就能够用于筛查化学结构数据库，并依据它们的互补性鉴定潜在的相互作用配体。这些算法试图利用空间约束和键能量方面的信息，使小分子与结合部位相适合。例子包括 AUTODOCK、LIGIN 和 GRAMM。化学数据库不仅能够通过一个结合部位（寻找互补的分子相互作用），还能通过另一个配体（寻找一致的分子相互作用）进行筛查。**合理药物设计**（rational drug design）这个过程已经引导了几个知名药物的生产，包括瑞乐沙和卡托普利。

19.5　小结

基因组计划产生大量的序列数据，必须应用计算机进行数据挖掘，以鉴定基因和调节元件。一旦可以得到基因，就必须确定它们的功能，以及各种基因产物间如何相互作用。在本章中，我们已经讨论了序列和结构分析如何用于提供一些关于蛋白质功能的信息，以及用于转录物组和蛋白质组分析的高通量技术的开发如何有助于将这些功能连接在一起。我们没有提到的功能基因组学的一个组成成分，是用于研究基因功能的突变体

的利用。因为这涉及细胞和动物的遗传修饰，所以我们将这个专题推迟到下一章，探讨基因操作如何用于研究基因功能和调节，以及建立人类疾病的模型。

（贺光　译）

进一步阅读

Eisenberg D, Marcotte EM, Xenarios I, Yeates TO (2000) Protein function in the post-genomic era. *Nature* **405**, 823–827.

Hanash S (2003) Disease proteomics. *Nature* **422**, 226–232.

Lee KH (2001) Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol.* **19**, 217–222.

Lockhart DJ, Winzeler (2000) Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836.

Pandey A, Mann M (2000) Proteomics to study genes and genomes. *Nature* **405**, 837–846.

Primrose SB, Twyman RM (2002) *Principles of Genome Analysis and Genomics*. Blackwell Science Oxford UK.

Various authors (1999) The Chipping Forecast. *Nature Genet.* **21**, (Suppl.) 1–60.

Various authors (2002) The Chipping Forecast II. *Nature Genet.* **32** (Suppl.) 465–551.

Vorm O, King A, Bennett KL, Leber T, Mann M (2000) Protein-interaction mapping for functional proteomics. In *Proteomics: A Trends Guide* 43–47.

参考文献

Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207.

Aicher L, Wahl D, Arce A, Grenet O, Steiner (1998) New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* **19**, 1998–2003.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.

Altman RB, Raychaudhuri S (2001) Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* **11**, 340–347.

Altschul SF, Madden, TL, Schaeffer AA, Zhang J, Zhang Z, Miller, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Aparicio S, Chapman J, Stupka E *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301–1310.

Audic S, Claverie J (1997) The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634.

Brenner SE (1999) Errors in genome annotation. *Trends Genet.* **15**, 132–133.

Brenner SE (2001) A tour of structural genomics. *Nature Rev. Genet.* **2**, 801–809.

Burton D (1995) Phage display. *Immunotechnology* **1**, 87–94.

Celis JE, Kruhoffer M, Gromova I, Frederiksenb C *et al.* (2000) Gene expression profiling: monitoring transcription and translation products using DNA microarrays and proteomics. *FEBS Lett* **480**, 2–16.

Der SD, Zhou A, Williams BR, Silverman RH (1998) Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays. *Proc. Natl Acad. Sci.* USA **95**, 15623–15628.

Dujon B (1996) The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270.

Eddy SR (1998) Multiple alignment and sequence searches. *Bioinformatics, A Trends Guide* **5**, 15–18.

Elgar G, Sandford R, Aparicio S, Macrae A, Vekatesh B, Brenner S (1996) Small is beautiful: comparative genomics with the pufferfish (*Fugu rubripes*). *Trends Genet.* **12**, 145–150.

Fields S, Sternglanz R (1994) The two hybrid system: an assay for protein-protein interactions. *Trends Genet.* **10**, 286–292.

Franzen B, Auer G, Alaiya AA, Eriksson E *et al.* (1996) Assessment of homogeneity in polypeptide expression in breast carcinomas shows widely variable expression in highly malignant tumors. *Int. J. Cancer* **69**, 408–414.

Gavin AC *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147.

Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29.

Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Research* **11**, 1425–1433.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Gorg A, Obermaier C, Boguth G, Harder A *et al.* (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis* **21**, 1037–1053.

Griffen TJ, Goodlet DR, Aebersold R (2001) Advances in proteome analysis by mass spectrometry. *Curr. Opin. Biotechnol.* **12**, 607–612.

Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999a) Quantitative analysis of complex protein mixtures using isotope coded affinity tags. *Nature Biotechnol.* **17**, 994–999.

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999b) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.* **19**, 1720–1730.

Hadley C, Jones D (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Struct. Fold Des.* **7**, 1099–1112.

Hardison RC (2000) Conserved non-coding sequences are reliable guides to regulatory elements.

Harrington CA, Rosenow C, Retief J (2000) Monitoring gene expression using DNA microarrays. *Curr. Opin. Microbiol.* **3**, 285–291.

Heinemann U, Illing G, Oschkinat H (2001) High-throughput three-dimensional protein structure determination. *Curr. Opin. Biotechnol.* **12**, 348–354.

Herbert BR, Harry JL, Packer NH *et al.* (2001) What place for polyacrylamide in proteomics? *Trends Biotechnol.* **19**, S3–S9.

Ho Y *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S, Sakaki Y (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147.

Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574.

Iyer VR, Eisen MB, Ross DT *et al.* (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87.

Jones DT (2000) Protein structure prediction in the postgenomic era. *Curr. Opin. Struct. Biol.* **10**, 371–379.

Khan J, Bittner ML, Saal LH, Teichmann U, Azorsa DO, Gooden GC, Pavan WJ, Trent JM, Meltzer PS (1999) cDNA microarrays detect activation of a myogenic transcription program by the PAX3-FKHR fusion oncogene. *Proc. Natl Acad. Sci. USA* **96**, 13264–13269.

Kolkman JA, Stemmer WPC (2001) Directed evolution of proteins by exon shuffling. *Nature Biotechnol.* **19**, 423–428.

Legrain P, Wojcik J, Gauthier JM (2001) Protein–protein interaction maps: a lead towards cellular functions. *Trends in Genetics* **17**, 346–352.

Lesney MS (2001) Pathways to the proteome: from 2DE to HPLC. *Modern Drug Discovery*, October issue, pp 33–39.

Liang P, Pardee AB (1992) Differnetial display analysis of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971.

Lilley KS, Razzaq A, Dupree P (2001) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. *Curr. Opin. Chem. Biol.* **6**, 46–50.

Lockhardt DJ, Dong H, Byrne MC *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680.

Mann M, Hendrickson RC, Pandey A (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437–473.

McCaffrey TA, Fu C, Du B, Eksinar S, Kent KC, Bush H Jr, Kreiger K, Rosengart T, Cybulsky MI, Silverman ES, Collins T (2000) Array-based screening identifies differential expression of heat shock proteins in human atherosclerotic lesions. *J. Clin. Invest.* **105**, 653–662.

Nadeau JH, Sankoff D (1998) Counting on comparative maps. *Trends Genet.* **14**, 495–501.

Norin M, Sundstrom M (2002) Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol.* **20**, 79–84.

Orengo CA, Jones DT, Thornton JM (2003) *Bioinformatics: Genes, proteins and computers.* BIOS Scientific Publishers, Oxford.

Orengo CA, Todd AE, Thornton JM (1999) From protein structure to function. *Curr. Opin. Struct. Biol.* **9**, 374–382.

Patton WF, Beecham JM (2001) Rainbow's end: the quest for multiplexed fluorescence quantitative analysis in proteomics. *Curr. Opin. Chem. Biol.* **6**, 63–69.

Pearl F, Orengo C (2002) Protein structure classifications. In: Orengo CA, Jones DT, Thornton JM (eds) *Bioinformatics: Genes Proteins and Computers.* BIOS Scientific Publishers, Oxford.

Phizicky EM, Fields S (1995) Protein–protein interactions: methods for detection and analysis. *Microbiol. Rev.* **59**, 94–123.

Quackenbush J (2001) Computational analysis of microarray data. *Nature Rev. Genet.* **2**, 418–427.

Rabilloud T (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but still it climbs up the mountains. *Proteomics* **2**, 3–10.

Schena M, Shalon D, Davis RW, Brown OP (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Sechi S, Oda Y (2003) Quantitative proteomics using mass spectrometry. *Curr. Opin. Chem. Biol.* **7**, 70–77.

Shapiro L, Scherer PE (1998) The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* **8**, 335–338.

Sillitoe I, Orengo C (2002) Protein structure comparison. In: Orengo CA, Jones DT, Thornton JM (eds) *Bioinformatics: Genes Proteins and Computers.* BIOS Scientific Publishers, Oxford UK.

Stulik J, Osterreicher J, Koupilova K, Knizek J *et al.* (1999) Protein abundance alterations in matched sets of macroscopically normal colon mucosa and colorectal carcinoma. *Electrophoresis* **20**, 1047–1054.

Suzuki H, Fukunishi Y, Kagawa I, Saito R *et al.* (2001) Protein-protein interaction panel using mouse full length cDNAs. *Genome Res.* **11**, 1758–1765.

Tucker CL, Gera JF, Uetz P (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.* **11**, 102–106.

Uetz P, Giot L, Cagney G, Mansfield TA *et al.* (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623–627.

Vasmatzis G, Essand M, Brinkmann U *et al.* (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl Acad. Sci. USA* **95**, 300–304.

Velculescu VE, Volgelstein J (2000) Analysing uncharted transcriptomes with SAGE. *Trends Genet.* **16**, 423–425.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. *Science* **270**, 484–487.

Velculescu VE, Zhang L, Zhou W, Vogelstein J *et al.* (1997) Characterization of the yeast transcriptome. *Cell* **88**, 243–251.

Venter JC, Adams MD, Myers EW *et al.* (2001) The sequence of the human genome. *Science* **291**, 1304–1351.

Wang H, Hanash S (2003) Multi-dimensional liquid phase based separations in proteomics. *J. Chromatography B.* **787**, 11–18.

Wasinger VC, Cordwell SJ, Cerpa-Poljak A *et al.* (1995) Progress with gene product mapping of the Mollicutes: *Mycoplasma genitalium. Electrophoresis* **16**, 1090–1094.

Werner T (2003) Promoters can contribute to the elucidation of protein function. *Trends Biotechnol.* **21**, 9–13.

Winter G, Griffiths AD, Hawkins RE, Hoogenboom HR (1994) Making antibodies by phage display technology. *Annu. Rev. Immunol.* **12**, 433–455.

Xenarios I, Eisenberg D (2001) Protein interaction databases. *Curr. Opin. Biotechnol.* **12**, 334–339.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, 15.

Zhang C, Kim SH (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.* **7**, 28–32.

Zhu H, Snyder M (2003) Protein chip technology. *Curr. Opin. Chem. Biol.* **7**, 55–63.

Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T (1998) Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. *Proc. Natl Acad. Sci.* **95**, 14470–14475.

Zhu H, Bilgin M, Bangham R, Hall D *et al.* (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105.

第20章　细胞和动物的遗传操作

本章内容

20.1　基因转移技术概述

培养的细胞和实验动物作为模型系统已经被广泛应用于健康和疾病的生物化学、生理学和发育等方面的研究。最近一段时间，这些研究非常得益于**基因转移技术**（gene transfer technology），基因转移技术能将特异性DNA序列导入动物细胞的基因组（**转基因发生**，transgenesis）。在基因转移技术问世以前，用于细胞和动物遗传改变的唯一方法是诱发突变，这种方法是用放射线或强化学试剂在固有的基因组内产生随机修饰。

基因转移技术的优点之一是它可以将体外制备新的DNA序列添加到基因组中。这些添加的序列可以是提供新功能的基因（功能获得），也可以是抑制内源基因表达的构件（功能丢失）。在另一种水平上，基因转移技术可以精确地或按预定的方式修饰固有的基因，因此，可以设计突变并将其引入体内特定的基因（**基因打靶**，gene target）。基因转移技术与复杂的基因表达调节方法的结合很大地提高了我们如何研究和控制基因功能以及构建小鼠疾病动物模型的能力。

将基因转移到培养的哺乳动物细胞中的首次报道是在20世纪60年代初期，当时发现人类的细胞可以从培养基中摄取和整合断裂的基因组DNA片段（Szybalska and Szybalski，1962）。虽然了解控制DNA转移的因素花费了一些时间，但在20世纪70年代初期，将DNA导入多种类型细胞就已成为常规方法。几乎在同时，诞生了第一个**转基因动物**（transgenic animal）。转基因动物的每一个细胞里都含有相同的附加的DNA序

列，并产生于将基因转移理论扩展到参与动物生殖系的细胞。第一个这样的动物是包括部分 SV40 DNA 序列（simian 病毒 40）的转基因小鼠，它是通过简单地将 DNA 注射到植入前胚胎的囊胚腔中制备的（Jaenisch and Mintz，1974）。此后，小鼠及很多其他物种包括无脊椎动物模式生物（果蝇和线虫）和鱼、蛙、鸟、大鼠及各种哺乳动物家畜等遗传修饰的可靠方法发展起来。最近，另外一种哺乳动物遗传操做的方法产生了很大的影响。1997 年当第一个成功的**哺乳动物克隆**（mammalian cloning）报道以后（Wilmut *et al.*，1997），遗传学进入了一个新的时代。它是将成体细胞的核移植到去核的卵母细胞中（**体细胞核移植**，somatic cell nuclear transfer），后来该技术作为另外一种常规方法被应用到遗传修饰动物的生产。

在本章的第一部分，我们描述了基因转移的原理和方法以及动物遗传操作的策略。在本章其余的部分和第 21 章，我们将讨论如何应用这些方法，重点是在生物医学研究领域的应用。

▶ **基因表达和功能的研究**（本章）。培养的细胞已经广泛应用于研究基因功能和基因表达的调节。这一应用的原因很简单：动物细胞为动物基因的研究提供了恰当的遗传背景和可靠的生物化学环境。用体外的实验系统精确地重新建立这样的背景是不可能的。在动物中添加基因、选择性删除和改变特异基因的能力为整个有机体内基因功能分析奠定了有力的基础。

▶ **在功能基因组学中的应用**（本章）。通过基因组插入突变文库、基因捕获文库和依据 RNA 干涉技术的高通量基因敲落技术的建立。基因功能可以被广泛地研究（节 20.2.6）。这些技术已经被应用到从秀丽新小杆线虫到小鼠等一系列模式动物以及培养的人细胞。他们补充了在第 19 章已讨论的各种功能基因组策略。

▶ **疾病模型的建立**（本章）。自然界已经提供了一些疾病动物模型，有些是通过随机突变产生的，但它们不是按照预想的方式产生的。用基因转移技术可以建立特定突变基因型动物，因此增加它们在遗传和表型的水平与人类疾病相似的机会。培养的细胞在非常有限的程度上也可用于疾病发生模型。

▶ **重组蛋白质的生产**（第 21 章）。培养的细胞和转基因动物已用于生产重组蛋白质的"生物反应器"。因为在哺乳动物系统中，重组蛋白质经历了真正的转录后修饰，因此它在生产治疗性人蛋白质上具有特殊的优势。

▶ **基因治疗潜力的开发**（第 21 章）。人体细胞的基因转移使基因缺陷的纠正和改善成为可能。

20.2　基因转移的原理

20.2.1　基因转移技术可用于将新的功能 DNA 序列瞬时或稳定地导入培养的动物细胞

基因转移技术常用于将新的功能基因添加到动物细胞中，即细胞获得表达一个或多个新的蛋白质的能力，这被称作**功能获得**（gain of function）。添加的 DNA 序列叫做转基因，尽管它有时可能含有一个、两个、甚至更多真正的基因；有时也叫"外来的 DNA"（foreign DNA），虽然这是一个误导，因为导入内源基因的外部拷贝也是完全可以接受的，并且这种方法对于基因功能的研究确实是非常有用的。另一个替代的词"外

源 DNA”（exogenous DNA）可被应用得更广泛和更可取。转基因可以用各种病毒和非病毒转移方法导入培养的细胞，这些方法归纳在框 20.1 中。

框 20.1　基因转移到培养的动物细胞中的方法

通常有四类基因转移方法可应用于培养的动物细胞，但转导和转染应用最为广泛。根据这些方法演变而来的许多方法都应用于在体内将 DNA 导入人类细胞（节 21.5）。

▶ **转导**（transduction）　这是病毒介导的基因转移，即外源 DNA 包裹在病毒颗粒内。动物病毒已经演化为可以通过各种方式将它们的 DNA 或 RNA 导入到细胞中，许多病毒因此被用于基因转移的载体，这些载体包括：

a）易于高水平瞬时转基因表达的病毒，例如腺病毒，新培斯（Sindbis）病毒，Semliki 病毒，痘病毒，杆状病毒（在昆虫细胞里该细胞支持杆状病毒复制）；

b）维持潜伏附着状态并有利于长期稳定转基因表达的病毒，例如埃巴（Epstein-Barr）病毒，单纯疱疹病毒，杆状病毒（在哺乳类细胞里该细胞不支持杆状病毒复制）；

c）整合到基因组用于持久稳定转化的病毒，例如反转录病毒，腺病毒相关病毒。

▶ **转染**（transfection）　用化学或物理的方法使细胞从培养基中吸收 DNA，DNA 通过一定的方式最后进入细胞核。需要注意的是在细菌系统，转染指的是吸收裸露的病毒（噬菌体）DNA，而转化是指吸收裸露的质粒或基因组 DNA。在动物细胞，转染指任何裸露 DNA 的吸收，而转化，如果用在基因转移中，通常指基因型稳定持久的改变。有很多不同的转染方法：

a）**化学转染**。当 DNA 与氯化钙混合溶解在磷酸盐缓冲液中，形成细小的 DNA-磷酸钙沉淀，此沉淀物沉降于培养的细胞膜上通过胞饮的方式被细胞吸收。DNA 也可以和各种其他化学试剂结合形成可溶的复合物，如 DEAE-葡聚糖（二乙胺基右旋糖酐）或去污剂 Polybrener。不适合磷酸钙覆盖的细胞用这些替代品效果更好；

b）**脂质体介导的转染**（图 21.8）。DNA 可被包裹在称做**脂质体**（liposome）的人造液体囊泡里，脂质体与细胞膜融合将 DNA 运送到细胞质中。病毒外膜通常含有促进膜融合的蛋白，因此用病毒外膜作脂质体可提高转染效率。这样的工具叫做病毒体。另外细胞膜也可以作为运输工具，这是原生质融合的基础，细菌原生质和 DNA 一起离心附着于哺乳类的细胞上，然后用聚乙二醇（PEG）诱导融合。相同的技术用于无血红蛋白的血影红细胞；

c）**脂质转染**。不像脂质体转染，DNA 被脂质囊泡包裹，脂质转染涉及 DNA-液体复合物（脂质复合物）的形成，它通过吞饮方式有效地被吸收。因此脂质转染与脂质体转染相比和化学转染有更多相同之处。最近，阳离子聚合体基因运输工具很受欢迎，它与脂质复合物一样有效但其优点在于特异性共聚体能够用于修饰运输工具的物理特性。在某些情况下，它可以在不同温度下形成具有不同物理特点的复合体，调控 DNA 的释放（Yokayama，2002）。这对于在体内将基因转移到身体特定位点非常有用（21 章）；

d）**电穿孔**。使用这种方法是将细胞暴露于短暂的脉冲电场中，使质膜产生瞬时的、纳米大小的孔道。如果 DNA 在缓冲液中的浓度足够，将通过这些小孔被吸收；

e）**受体介导的内吞**（图 21.9）。在这种方法中，DNA 附着在细胞表面受体的配体上，受体通过吞饮方式被回收，DNA 在处理受体的过程中被释放。在正常环境下，包含受体-配体-DNA 复合体的核内体融合到溶酶体中，而 DNA 被降解，但因为这些破坏核内体，可通过在转染复合体上含有腺病毒多肽而被阻止。这种方法能根据它们受体的特异性来靶定特异细胞类型。

框 20.1　基因转移到培养的动物细胞中的方法（续）

- **直接转移**　这种方法是直接通过物理方法将 DNA 导入细胞。最典型的例子是**显微注射**（microinjection），偶尔用于其他基因转染方法无效的培养细胞，但通常用于动物的卵、配子或早期胚胎。**颗粒轰击**（particle bombardment），是将表面附有显微投射弹的高速 DNA 加速注射到细胞内，是另外一种直接转移方法。这种方法可用于培养的细胞，但它是选用于组织切片细胞的转染，同时也可用于体内基因转移（21.5.4 部分）。
- **细菌基因转移**　典型的这种方法使用活的侵袭细菌，细菌在动物细胞内溶解释放其装载的 DNA（Higgins and Portnoy，1998）。沙门氏菌种溶解发生在噬菌细胞的囊泡里，而其他菌种［例如单核细胞增生利斯特氏菌（*Listeria monocytogenes*）and 弗氏志贺氏菌（*Shigella flexneri*）］溶解发生在细菌逃离囊泡后。细菌也可以附在细胞表面通过鞭毛转移 DNA。根癌土壤杆菌（*agrobacterium tumefaciens*）用这种方法将 DNA 转移到植物中，也能感染培养的动物细胞（Kunik *et al.*，2001）。

转基因的结局很重要，有的情况在宿主细胞中瞬时维持，而在另外的情况它成为基因组永久的一部分。如果用病毒介导的基因转移，转基因的结局依赖于病毒的特点，一些病毒只适用于瞬时表达，而有些则能够潜伏感染，可使转基因长期表达。反转录病毒特别之处在于其 DNA 拷贝在感染后不久能整合到宿主基因组。因此重组的反逆转录病毒可用于建立稳定的转化细胞系。

当 DNA 通过转染被导入培养的动物细胞时，通常是以细菌质粒的形式，该质粒不能在动物细胞内复制，大部分细胞开始可以摄取 DNA，但不久就被降解，任何转基因表达都是瞬时的。质粒的寿命依赖于 DNA 的质量和宿主细胞系。在大多数细胞内，高质量的质粒 DNA 可持续 1～2 天，但有时可持续更长时间（例如，在 HEK293 细胞系，质粒 DNA 可持续达 80h）。在很少的转化的细胞中，一些质粒 DNA 整合到基因组，导致**稳定转化**（stable transformation）。这种整合事件很罕见，因此必须用有效的筛选策略来确定稳定转化的细胞。一般的策略是在质粒上包含一个**可筛选的标记基因**（selectable marker gene），或者用两个质粒共转染细胞，一个质粒含有要转移的基因，一个质粒含有筛选标记。可筛选的标记赋予细胞一种特性即让它在有某一**筛选试剂**（selective agent）时存活，例如一种抗生素，可杀死非转化的细胞（框 20.2）。用这种方法能够建立稳定的转化细胞系。

框 20.2　动物细胞的可筛选标记

可筛选的标记基因的种类

一个可供筛选的标记基因应该具有一种特性，即当存在某些特殊试剂时非转化细胞死亡或生长受到限制，而稳定转化细胞能够生存并生长。因为细胞是在携带标记基因的基础上被筛选的，所以叫做**阳性筛选**（positive selection）（对于标记基因来说是阳性的）。阳性筛选对于在培养皿中分离用非复制 DNA 转染后稳定转化的少数细胞是非常必要的。有两大类可供筛选的标记基因：内源性可筛选标记和显性可筛选标记。

- **内源性可筛选的标记**　标记基因存在于宿主细胞基因组内，因此选择只适用于缺少该基因功能拷贝的细胞。如 *TK* 基因编码**胸苷激酶**（thymidine kinase）。这个酶可将游离的胸腺嘧啶转换

框 20.2　动物细胞的可筛选标记（续）

为磷酸胸腺嘧啶，它是**核酸代谢旁路**（nucleoside salvage pathway）的一部分。在正常的环境，细胞不需要此旁路代谢，因为胸腺嘧啶脱氧核苷酸可以由磷酸尿嘧啶从头合成。但是如果开始合成途径被抑制剂**氨基蝶呤**（aminopterin）阻断，此细胞就变成依赖 TK 活性。因此 TK^- 的细胞（缺乏 *TK* 基因功能）只有稳定转化 *TK* 基因后才可以在 HAT 培养基中（包含氨基喋呤和胸腺嘧啶）生存。

▶ **显性可筛选的标记**　标记基因是宿主细胞基因组内没有的基因，并赋予一种全新的特性诸如抗生素耐药性。这种标记的好处是可用于任何类型的细胞，即不需要突变体。例如，大肠杆菌的新霉素基因编码新霉素磷酸转移酶。它能使氨基甙类抗生素如 G418 失活，在培养基中加入 G418 就可以很简单地筛选到稳定的转化细胞。

可扩增的标记基因

如果选择的筛选试剂是标记基因产物的竞争性抑制剂，那么可能要用**逐步地筛选**（stepwise selection）来扩增标记基因，并获得高水平转基因表达。用此系统进行筛选是因为在稳定的转化细胞中转基因座通常含有可筛选标记的多个串联拷贝和最初的转基因。通过增加筛选试剂的浓度来筛选高拷贝数的细胞，因为这些细胞可高水平表达标记基因。在这些细胞中可发生事件重组，进一步增加拷贝数和选择压力而逐渐地选择具有大量扩增转基因序列的细胞。扩增是随机的过程，最初的转基因伴随标记基因同时扩增。这种标记的例子就是小鼠 *dhfr* 基因，它编码**二氢叶酸还原酶**（dihydrofolate reductase）。可用竞争性抑制剂**氨甲喋呤**（methotrexate）逐步选择来扩增。

反向可筛选的标记基因

一个**反相可筛选标记**（counter selectable marker）提供一种特性即能杀死稳定的转化细胞而允许非转化细胞存活。这称为阴性选择，是因为缺少标记基因（它们对于标记基因来说是阴性的）的细胞被筛选。如果这样标记在控制限制性启动子的情况下表达，即它们对细胞消除是有用的。对于确认特殊类型的基因修饰也是有用的，例如在 ES 细胞中用来区别基因打靶事件和随机整合事件（节 20.2.4）。一些反向标记可直接杀死细胞（如蓖麻蛋白毒素，白喉毒素等），而其他的需要筛选试剂[如 TK 激酶可用于胸苷酸类似物 9-（1，3-二羟-2-丙氧基鸟嘌呤）存在时的阴性筛选]。

除此之外，可以用含有病毒复制起始点的质粒载体转染细胞，这样转基因被维持并异位复制。有些载体，起始点使质粒复制快速并失去控制，导致转基因表达水平瞬间增高，但后来杀死细胞。用携有 SV40 病毒复制起始点转染 COS 细胞就是这种情况；相反，携带延迟复制起始点埃巴病毒（EBV）的载体可维持在中等拷贝数，如果带有适当的筛选标记可用于建立多种稳定的转化细胞系。

总之，病毒和非病毒基因转移方法都可用于把新的基因瞬时或稳定的导入动物细胞。可通过反转录病毒整合，复制载体长期易位维持或非复制 DNA 整合到一条宿主细胞染色体等方法获得稳定的转化。

20.2.2　转基因动物的制备需要将基因稳定地转移到生殖系

以上讨论的方法适用于在培养皿中的细胞转化，但将这种方法拓展到整个动物上是非常困难的。我们不可能将一个成体动物的所有细胞都同样地转入 DNA。因此，为了制备一个完全的转基因动物（一个动物的所有细胞都带有相同背景的同样的转基因），这个动物必须由一个**转基因的合子**（transgenic zygote）发育而来。这要通过基因稳定地转移到生殖系而获得，并可通过以下二种途径之一来实现（图 20.1）。

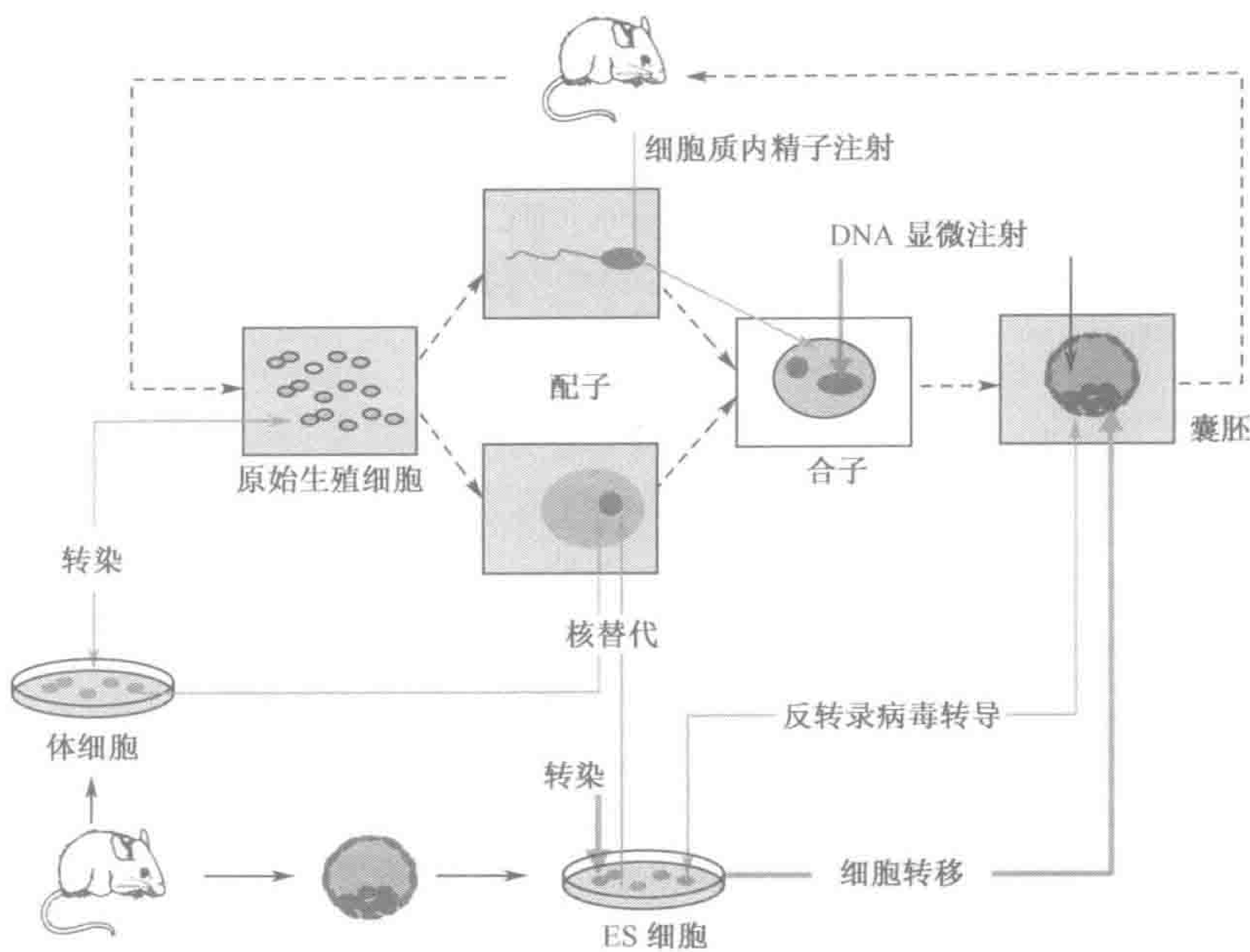


图 20.1　制作遗传修饰小鼠的多种途径

用点线连成的框显示了小鼠的生命周期，代表了能进行遗传修饰的所有阶段：生殖细胞、配子、合子、囊胚和成体。图下面的部分显示了另外一个用于核移植供体细胞来源的小鼠和细胞移植过程。红色箭头显示了外源 DNA 的输入。粗箭头代表了广泛用于小鼠基因移植的方法。

- 直接地，通过将 DNA 选择性地导入生殖细胞，由此衍生配子细胞或导入刚刚受精后的卵细胞。如果 DNA 在合子第一次分裂之前整合，由此发育而来的动物的每一个细胞都将含有相同的转基因。
- 间接地，通过非选择地将 DNA 导入生殖系形成前的胚胎。在这个阶段，胚胎细胞是全能的，或者至少是多能的，这意味着它们在胚胎发育中可以形成任何类型的细胞。由于 DNA 是在合子第一次分裂之后发生整合的，因此只有部分胚胎细胞会整合转基因。然而，如果这些细胞参与胚胎的生殖系，将产生转基因的配子，其后代将是转基因动物。

表 20.1 概括了基因转移到动物的方法，并在下面进行更详细讨论。

表 20.1　动物基因转移：打靶和转化生殖系方法概述

靶细胞	方法
生殖细胞	转染培养的原始生殖细胞(哺乳动物)
	注入生殖细胞发育阶段的胚胎(果蝇)
精子	精子头吸附 DNA(哺乳动物)
	DNA 导入去浓缩的精子细胞核内(非洲爪蟾)
卵/合子	显微注射到卵细胞质(鸟、两栖类、鱼、秀丽新小杆线虫)
	前核显微注射(哺乳动物)
	反转录病毒转移(哺乳动物、灵长类)

续表

靶细胞	方法
	核移植
囊胚	囊胚腔显微注射(哺乳动物)
	ES 细胞转移(小鼠)
	反转录病毒转移(哺乳动物、鸟)
ES 细胞(接着细胞转移)	转染→转基因添加
	转染→基因打靶
	反转录病毒转移
体细胞(接着核移植)	转染→转基因添加
	转染→基因打靶
	反转录病毒转移

直接将基因转移到生殖系前体细胞是制备转基因果蝇的标准方法

DNA 导入生殖细胞可能是一种非常有用的进入生殖系的方法，因为通过这种方法可以产生转基因配子。这种方法并没应用到哺乳动物当中是因为：虽然哺乳动物的原始生殖细胞相对容易分离、培养和转染，但修饰后的生殖细胞重新导入宿主动物后很难再形成生殖系细胞。但生殖细胞的直接修饰是制备转基因果蝇的常规方法。在黑腹果蝇，用称为 P-因子（P-element）的可转座元件序列获得了有效的染色体整合。转基因被插入到两个 P-因子的末端序列之间，然后注射到有极性细胞（生殖系的前体细胞）核的幼龄胚胎的极性胞质中。P-因子转座需要的转座酶是由共同注射的质粒提供的，这种方法有利于转基因随机整合到一个或更多的极体细胞核的基因组中。通常只整合一个转基因的拷贝。

DNA 能与精子或精子的核混合并导入未受精的卵细胞

配子的直接修饰是制备转基因动物的另外一种方法。目前发明了两种非常不同的方法：精子介导的 DNA 转移和限制性酶介导的整合。

精子介导的 DNA 转移（sperm-mediated DNA delivery）是利用精子头在体外自发地与 DNA 结合的事实。因此，虽然配子基因组是未经修饰的，但精子可以用作传送工具。**ICSI（细胞质内精子注射**，*intracytoplasmic sperm injection*）是一种不孕的治疗方法，它是将精子头注射到卵细胞的细胞质里。*ICSI* 方法已经应用于将质粒 *DNA* 包裹的精子头注射到哺乳动物的卵中。在最早的这种实验中，将带有绿色荧光蛋白（*GFP*）基因的质粒附着在小鼠的精子上，然后注射到分离的卵细胞中。几乎所有被注射的卵都显示了绿色荧光蛋白的活性，但只有 20% 产生了转基因鼠，这表明在大多数情况，转基因没有整合到细胞的基因组（*Perry et al.*，1999）。同样的技术已经应用于恒河猴的实验，但是尽管有几个胚胎显示了瞬时 GFP 活性，却没有得到一个转基因猴。

限制性酶介导的整合（restriction enzyme-mediated integration，REMI）是建立制备转基因蛙的方法。不像精子介导的基因转移，配子基因组是在基因转移的过程中修饰的。精子的核被分离、去浓缩与质粒 DNA 混合。然后用一定量的限制性酶处理，使精子的核产生缺口。然后去浓缩的核被移植到未受精的卵细胞中，核的缺口被修复，导致

质粒 DNA 整合到基因组。核是非常脆弱的，因此移植过程必须快速完成（Kroll and Amaya，1996）。

显微注射 DNA 到受精卵中是已确立的制备转基因小鼠的方法，并用于鱼和两栖类的瞬时表达检测

通过将 DNA 注射到受精卵的细胞质中已经产生了转基因小鼠，但是一个更有效的技术亦已建立，即将 DNA 直接导入刚刚受精后的雄原核。图 20.2 显示了这种方法。显微注射的转基因随机整合到染色体的 DNA，通常是在一个单一位点，并以多个拷贝整合（50 或以上的首尾相连的拷贝很常见）。转基因可迅速发生整合，在这种情况下产生的小鼠就是转基因的。然而，DNA 在 1 个细胞或 2 个细胞分裂后发生整合更为常见，这种情况，产生的小鼠是一个**嵌合体**（mosaic），它既含有转化细胞也含有非转化细胞。如果转化的细胞发育成生殖系细胞，那么转基因就传递给下一代小鼠，并可通过 PCR，Southern 印迹或转基因表达检测等予以确证。显微注射的小鼠卵 40%可获得生殖系的传递，遗憾的是当这种方法被应用到其他哺乳动物时传递率很低（<1%），这部分原因是由于卵的处理困难，部分是由于卵的生存率低。

显微注射也可用于将 DNA 导入鱼和两栖类卵，但与哺乳动物的情况不同，该

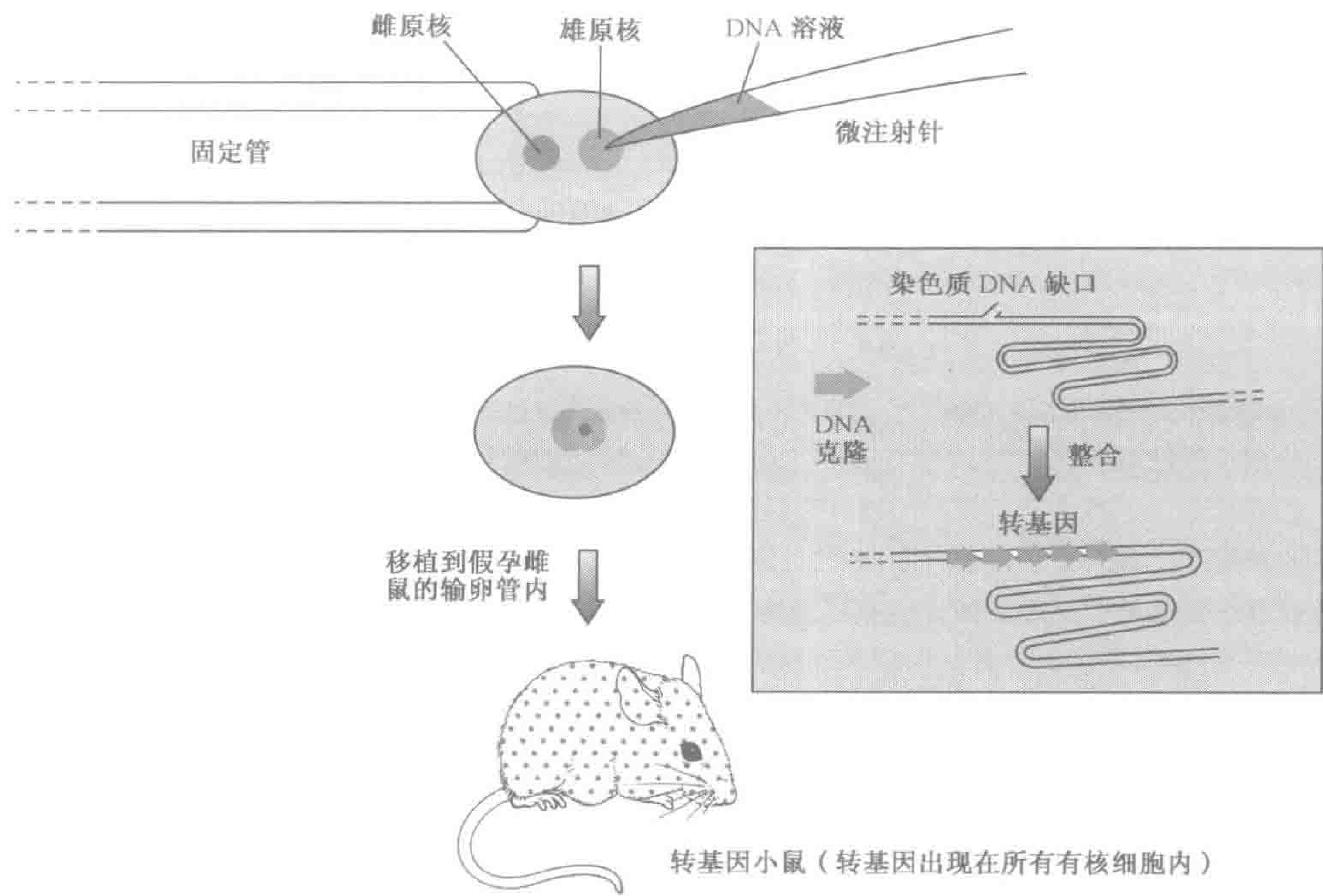


图 20.2　通过原核微注射构建转基因小鼠

非常纤细的玻璃管是用针特殊的设备制作的：一是固定管，有一个能容纳部分受精卵的管口，并使受精卵固定于此，而微注射针有一个很细的尖，用于刺入受精卵，再进入雄原核（因为雄原核较大）。每份所需的 DNA 液体直接注入到原核中，导入的 DNA 克隆能在缺口处整合到染色体 DNA 上，形成转基因，通常含有多个头尾相接的拷贝。微注射针退出后，活的卵细胞植入到假孕雌鼠的输卵管内（和结扎雄鼠交配；能启动刺激植入胚胎发育的雌性生理改变。）由移植胚胎发育成的新生小鼠可用 PCR 方法检测是否含有目的 DNA 序列。

DNA 长期存在趋向异位状态并广泛复制。注射的质粒 DNA 在发育的早期可增加 50～100 倍。这反映了在鱼和两栖类发育的早期没有转录的事实，所以广泛储存了 DNA 复制所需要的酶和蛋白质。因此鱼和青蛙的卵可用于瞬时表达系统。然而，一些 DNA 发生了整合，如果建立了生殖系传递就可获得转基因系。这种方法不适用于世代间隔长的蛙，但它是制作转基因鱼的标准方法。

反转录病毒基因转移已用于制备转基因哺乳动物和鸟，但此法主要用于建立发育的嵌合体

基因可用反转录病毒转移到未经筛选的早期胚胎细胞，因为感染后一个载体 DNA 拷贝会稳定整合到宿主基因组内。用重组小鼠反转录病毒感染植入前小鼠胚胎或将反转录病毒注射到早期植入后小鼠胚胎可导致一些细胞的稳定转化，并因此产生嵌合体，这些嵌合体可繁育出转基因后代。用鸟类反转录病毒已在鸡获得相似的结果。但是，由于反转录病毒基因转移的一些限制（能够结合的外源 DNA 量的限制和反转录病毒转基因沉默的趋势）这一技术没有广泛的用于制备转基因小鼠或鸟。反之，它有利于鱼的嵌合体制备，可用于脊椎动物发育中的基因表达和功能的研究。最近，将重组反转录病毒注射到分离后的卵细胞的卵黄周围腔隙中已制备了转基因牛（Chan *et al.*，1998）和第一个转基因灵长类动物——一个命名为 ADNi 的恒河猴（Chan *et al.*，2001）。

通过胚胎干细胞转染制备转基因鼠

小鼠胚胎干（ES）细胞来自于交配后的 3.5～4.5 天的胚胎，且来源于囊胚的内细胞团（框 20.3）。ES 细胞可以在体外培养，并且用框 20.1 中介绍的方法很容易转染。

框 20.3 哺乳动物胚胎干细胞的分离和操作

在哺乳动物，正常的胚胎来源于囊胚的内细胞团（ICM）（节 3.7.3）。小鼠胚胎干（ES）细胞第一次是由 Evans 和 Kaufman（1981）及 Martin（1981）分离的。其过程是将 4.5 天植入前胚胎（囊胚）放在单层的饲养细胞上，饲养细胞提供一个附着基质并能分泌抑制 ES 细胞分化的蛋白质因子。囊胚附着后，内细胞团开始增殖。在适当的时间，用微量移液器将内细胞团自然地移出使其分散成小的细胞团接种在新的饲养层细胞上。在显微镜下检查克隆的形态学特点。然后挑出并使其分散成单个细胞重新接种在新的饲养层细胞上。最后分离具有相同形态的细胞，并能建立细胞系。引入“ES 细胞”这个词是为了区别来自胚胎的干细胞和来源于畸胎瘤的癌胚细胞（EC）。通常 ES 细胞比 EC 细胞具有更高的发育潜力。这两类细胞都是多能的，而 ES 细胞可分化成所有类型的成体细胞，包括生殖系细胞。然而它们不是像受精卵意义上的全能细胞：如果将 ES 细胞植入子宫，他们不能发育成胚胎。有报道 EC 细胞在一些情况下可发育成嵌合体的生殖系，但将小鼠 ES 细胞注射到从不同品系分离的囊胚中并植入到假孕母鼠体内可形成嵌合鼠，尤其是发育成更为一致形式的生殖系。就是这个特点使小鼠胚胎干细胞在研究中如此珍贵。但应该注意到，ES 细胞用于成功地形成生殖系嵌合体是来自专一小鼠品系（129）与专一宿主胚胎品系（C57BL/6）的结合。ES 细胞在这种组合中生长旺盛并容易发育成生殖系。研究者对其他哺乳动物 ES 细胞的研究已有多年，虽然从小鼠以外的其他动物中已经分离得到了 ES 细胞，但并没有得到与小鼠生殖系嵌合体平行的巨大成果。最近，人类的胚胎干细胞已经从囊胚（Thomson *et al.*，1998）和原始生殖系（Shamblott *et al.*，1998）中分离。人类 ES 细胞潜在的医学应用及其应用的伦理影响将在 21 章讨论。

然而这些细胞仍是多能的，并参与小鼠的各种组织——包括生殖系的形成。当把它们注射到一个宿主囊胚中，并将其再植入假孕的代养母鼠体内，这个胚胎将发育成**嵌合体**(chimera)。该嵌合体包含来源于不同合子的两个细胞群，即来自囊胚的细胞和植入的 ES 细胞。这不同于细胞遗传上可能不同但来源于同一个合子的嵌合体（图 4.10）。如果囊胚和 ES 细胞来源于不同毛色的小鼠，嵌合鼠的后代很容易通过它们镶嵌的毛色进行判断。转基因鼠的生殖系传递也可以通过嵌合鼠（通常为雄性）与 ES 细胞源性但毛色隐性的雌鼠进行交配、筛选来证实（图 20.3）。

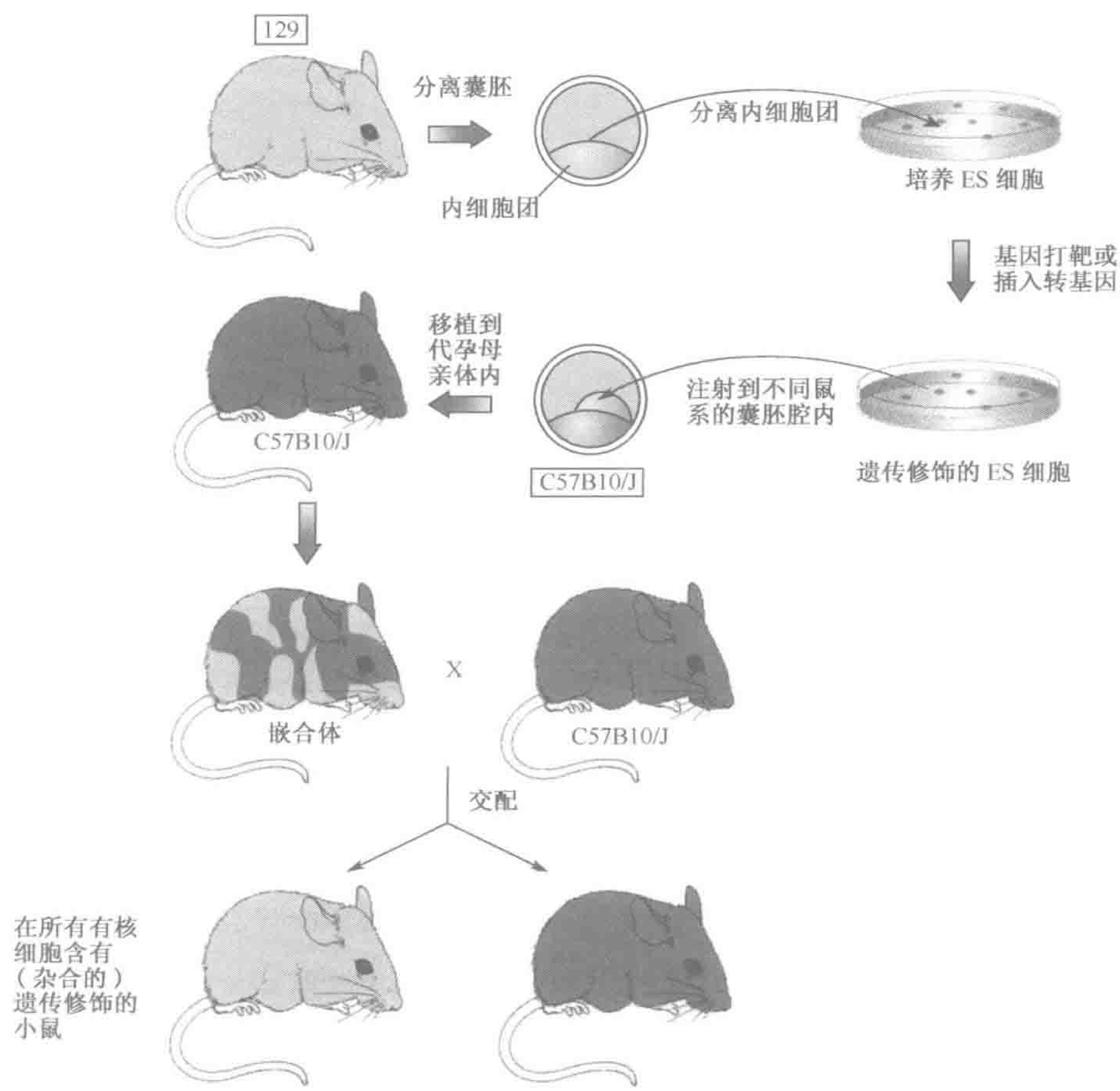


图 20.3　用于外源 DNA 或特定突变转移到小鼠生殖系的遗传修饰的 ES 细胞

从一个稳定的小鼠品系（129）切除其输卵管或分离目的囊胚内细胞团的细胞进行培养，这样的胚胎干(ES) 细胞保持了可以最终分化成成体小鼠不同类型组织的能力。ES 细胞在培养阶段可通过外源 DNA 的插入或导入精细突变进行遗传修饰。然后被修饰的 ES 细胞被注射到另外一种品系小鼠（如 C57B10/J，被毛是黑色的，对 129 品系小鼠被毛是隐性的）分离的囊胚内，然后移植到和囊胚同一品系的假孕母鼠体内，随后被诱导的囊胚发育成包含两类不同来源的细胞群（包括生殖系细胞）的**嵌合体**（chimera）（正常情况下具有镶嵌性毛色），嵌合体回交后产生的小鼠其遗传修饰是杂合的，随后杂合突变体的互交可产生纯合子。

ES 细胞的优点之一是它可以在体外无限地进行培养，因此它有培养细胞所具有的所有优点。我们可以用简单的转染技术转化大量的细胞，并且可以在培养阶段通过可筛

选标记诸如新霉素来验证想要的遗传修饰（框 20.2）。相比之下，上述讨论的其他方法需要对单个卵和胚胎进行手工操作，所以只能做少量实验。而且也没有筛选转化的卵或胚胎的方法，因此转基因整合必须在产生的小鼠中进行确认。然而，ES 细胞最大的优点是它们易于发生同源重组，允许通过基因打靶进行遗传修饰（节 20.2.5）。近来，已获得鸡和人的 ES 细胞系（见框 20.3 和节 21.3.3 的讨论），但从家畜中获得各种健康可靠的 ES 细胞系还远不可能。

核移植可用于家畜的遗传修饰但成功率很低

由于转基因鼠制备的主要技术（原核显微注射和 ES 细胞转染）在大多数其他哺乳

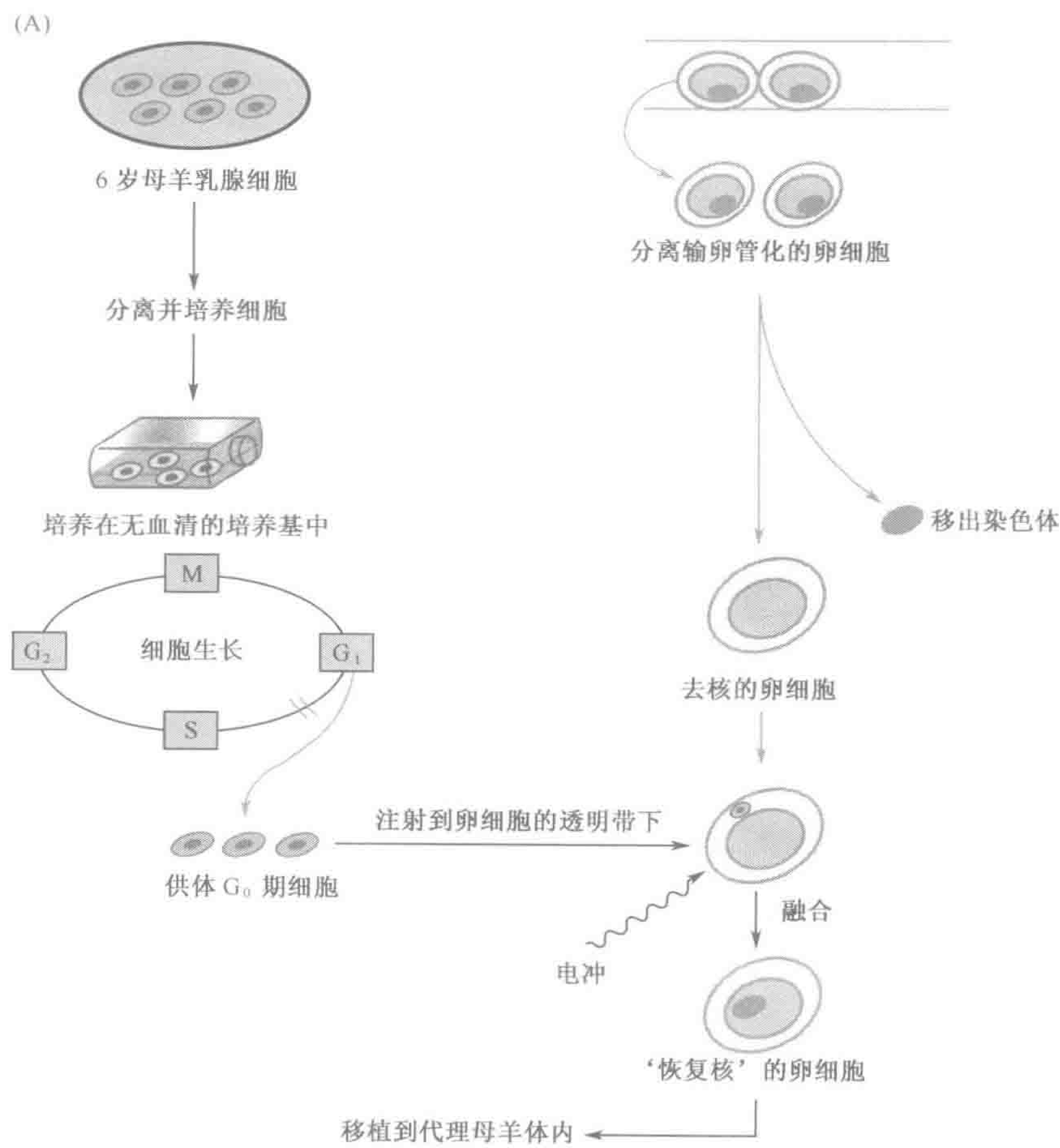


图 20.4　多利羊是由成体细胞尝试哺乳动物克隆的第一个成功的成果

(A) Wilmut 等（1997）使用的实验策略。供体细胞核来自于成体乳腺细胞建立的细胞系。核转移是通过将单个的体细胞与中期Ⅱ去核卵母细胞融合而完成的。供体细胞在使用前去除血清，迫使细胞周期停滞进入静止状态的 G_0 期，这一时期只发生少量转录。由于正常情况下卵细胞与转录失活的精子受精，精子的核可能被卵细胞内的转录因子或其他有用的染色质蛋白程序化，此 G_0 期的核代表了程序化这种理想的基本的状态。需要注意的是在其他的克隆实验中，使用了不同的策略将供体核导入卵细胞。例如，在 Wakayama 等（1998）成功克隆成体小鼠中，用一个很细的针吸出供体细胞核携带少量供体细胞的细胞质。供体细胞快速但轻柔地注射到去核的卵中。(B) 多利和它的第一个孩子鲍尼，Roslin 学院友好馈赠的原始照片。

(B)



图 20.4　(续)

动物是无效或不可行的，因此，开发了以**核移植技术**（nuclear transfer technology）为基础的替代方法。核移植涉及用体细胞核替代卵细胞核，然后体细胞核被卵细胞重新程序化，无论供体细胞的分化状态如何，使其能够重演发育的整个过程。

该技术本身不是一项新的技术，它用于产生克隆的两栖类已有 50 多年，自从 20 世纪 80 年代末期已用该方法从胚胎细胞产生克隆的哺乳动物。1995 年第一次表明利用培养细胞的核能克隆哺乳动物（Campbell *et al*.，1996），并于 1997 年第一次从成体细胞产生了克隆的哺乳动物（Wilmut *et al*.，1997）。图 20.4（A）中概要介绍了 Wilmut 及其同事们使用的技术流程。在最初的实验中，434 个卵细胞中只有 29 个发育到可转移阶段，并且这 29 个卵细胞中只有一个发育到最后阶段即闻名的“多利”羊（图 20.4B）。成体动物克隆的成功迫使我们接受了曾经以为不可逆转的基因组修饰是可逆转的这个事实，并且成体细胞的基因组可以在卵细胞中受一些因子的作用重新程序化再一次变为全能细胞。深入对发育过程中的基因表达调控和体细胞分化、突变、衰老以及修复过程等研究将毫无疑问得益于动物克隆，尤其是小鼠的克隆（Wakayama *et al*.，1998）。我们在第 21 章讨论了克隆鼠和克隆人实际的医学价值和伦理影响，但这里我们只把这种方法看作是另外一种制备转基因动物的途径。重要的一点是，如果供体细胞核是经过转染的，有添加的转基因，那么由这个被操作的卵细胞发育而来的动物就将是转基因动物。第一次证实这种设想的是 Schnieke 等（1997），他用人的凝血因子 IX 基因转染培养的羊成纤维细胞生产一个克隆的转基因羊（Polly），在它的羊奶中产出了重组蛋白。通过同样的方法从培养的用基因打靶改变了基因组的体细胞获得了克隆羊（见下文和 McCreath *et al*.，2003）。

20.2.3 转基因表达的调控是任何基因转移实验都需要考虑的重要因素

在一个细胞或转基因动物中，转基因的出现其自身并不足以能产生功能的产物，要发生这种情况转基因必须是表达的。因此基因表达调控在基因转移技术中是至关重要的。转基因表达是由表达构件上的序列调节的，即使当转基因整合了，它也会受宿主基因组内一些因素的调节。总之，启动子的结构是构件设计的最重要方面，也是下面我们集中讨论的问题。然而正常表达还需要多聚腺苷酸化位点，可能还有其他需要考虑的因素，诸如适合的翻译起始位点和含有合适的靶蛋白信号等。

转基因启动子确定了表达的基本时空模式

在细胞系的基因转移实验常得益于非常活跃的**组成性启动子**（constitutive promoter）的使用，它能使转基因在任何时候都能很强地表达。这样的启动子一般来源于病毒，它们已经演化成在许多不同类型的细胞中都表达它们的基因。哺乳动物细胞中最常用的调节元件，包括 SV40 早期启动子和增强子，Rous 肉瘤病毒长末端重复序列（LTR）启动子和增强子以及人巨细胞病毒介导的早期快速启动子。在许多商品化的载体中都包含了这些启动子。

在转基因动物中，经常需要在特定组织或特定发育阶段表达转基因。同样，在细胞系，对细胞周期特定阶段需要限制转基因表达或只在细胞分化阶段开启转基因表达。通过将转基因与合适的细胞-或阶段-特异性启动子连接在一起，就可获得所期望的表达模式。例如，神经特异性烯醇化酶基因只有在成熟的神经元中表达，其上游 1.8kb 长的调节元件足以调节转基因小鼠中的神经组织特异性表达，并在分化培养的神经纤维细胞中特异性地使转基因表达上调（Forss-Petter *et al.*，1990；Sakimura *et al.*，1995）。

在细胞系和动物基因通过**诱导型启动子**（inducible promoter）最大限度地调控基因表达，它可通过调控一个特殊化学配体的供给而开启或关闭基因。典型的调节这样诱导型启动子的转录因子是通过这个配体进行结构修饰的。有几个自然的诱导型启动子已被使用，包括小鼠金属蛋白酶启动子和小鼠乳腺癌病毒 LTR 启动子，这两个启动子都是受地塞米松（一种合成的激素）诱导的。另外，金属蛋白酶启动子是受重金属离子诱导，例如 Zn^{2+} 诱导配体可被添加到细胞培养基中或补充到饲养动物的饮水中。一般来说，这些内源性启动子的使用受“泄露”影响，即高背景表达和相对低水平诱导表达。对相同的配体有反应的内源基因共活化可能会产生不良的后果，并且就转基因动物来说在不同的器官对配体的吸收或降解效率可能不同。然而，最近通过用异源成分开发了很有前景的诱导系统，例如，大肠杆菌（*E. coli*）*tet* 操纵子，它是四环素调节诱导表达系统的基础（图 20.5 A），该系统允许产生高效可诱导的转化细胞系和转基因动物（Saez *et al.*，1997）。根据大肠杆菌乳糖操纵子（*lac*）和果蝇蜕皮激素还设计了其他的系统。另外，诱导表达也可以通过**化学诱导的二聚体**（chemically-induced dimerization，CID）来完成，它是由两个二价配体装配的功能转录因子（Belshaw *et al.*，1996）。这些表达系统的一个弊端是诱导发生在转录水平，因此在诱导的反应之前可能明显延迟，以及在诱导刺激物除去后和恢复基础水平状态之前也有同样的延迟。如果快速诱导和延迟都很重要，可使用在蛋白水平工作的诱导系统，在这样一个系统中靶基因

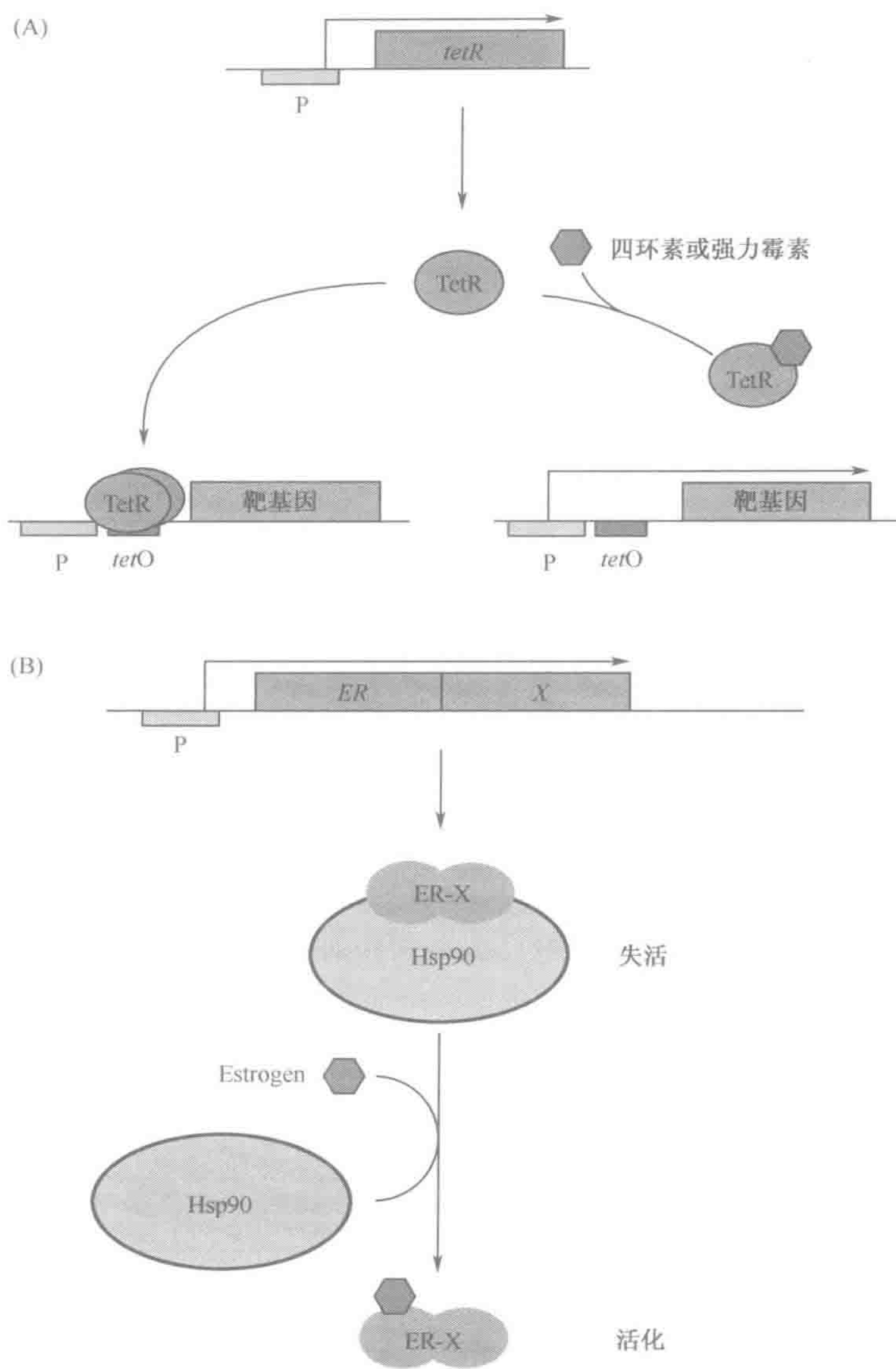


图 20.5　诱导表达系统

(A) 四环素诱导基本表达系统 *E. coli* 四环素抑制子（TetR）的组成型表达将使任何含有与抑制子结合的四环素 *Tet*O 操纵子序列的打靶基因失活，然而当有四环素或四环素类似物强力霉素存在的时候，抑制子的构象发生改变，不再与操纵子结合，引起基因去抑制。值得注意的是抑制子高水平组成型表达的毒素效应限制了这种系统在一些细胞中的应用。为了解决这个问题，发展了更复杂的系统，在这个系统中四环素抑制子能转变为一个激活子（tTA）或者结合依赖于四环素而不是通过四环素释放（反向 tTA 系统）。见 Saez 等（1997）关于这一系统的综述。(B) 雌激素诱导的表达系统。蛋白质（X）与雌激素受体（ER）融合表达，通常情况由于它与热休克蛋白 90（HSP90）结合形成复合物而失活。然而，在雌激素存在时融合蛋白从复合物上释放，蛋白 X 活性恢复。这是一个很好的系统，因为诱导很快，只需要从复合体中释放，而不需要随后的蛋白合成的转录。

表达产物与雌激素受体融合，通常情况下雌激素受体与 Hsp90 结合被隔离为失活复合

体（图 20.5B），当雌激素或其类似物三苯氧胺存在时雌激素受体被释放，而与之融合的蛋白质能被激活（Littlewood *et al.*，1995）。

整合的转基因表达受位置效应和基因座结构的影响

通常观察到，独立获得的带有相同转基因构件的转基因动物并不总显示其转基因表达相同的水平或模式。这是因为转基因整合的发生是随机的，因此转基因座的位置和结构都是可变的。位置效应是受局部调节元件和染色质结构的影响而产生的。例如，转基因可能整合到一个增强子附近，其表达模式受此增强子的修饰；或者转基因也可整合到异染色质结构域内，其表达一起被抑制。转基因座的结构也影响转基因表达。例如，如果转基因的两个拷贝恰好反向重复排列，则产生发夹样 RNA，导致 RNA 干涉（节 20.2.6）。还有基因座结构各种其他方面可触发细胞对入侵 DNA 的防御，通过 DNA 从头甲基化导致基因沉默。这些现象在转化的细胞系不太明显，因为这是根据它们高表达标记基因的能力筛选出来的。

位置效应可通过采用显性作用调节元件和大的转基因构件来阻止

大多数转基因是由短的调节元件来控制 cDNA 序列。很明显，在确定基因表达和功能所需的最小序列的这样元件时并没有在基因组背景上提供基因高效表达所需的全部序列。有证据表明，特定的调节序列作为对建立开放染色质结构域的控制开关和保护基因不受相邻结构域内调节元件和染色质结构的影响。这些元件经常存在于内含子和远离的位点。因此，位置效应可通过整合这些调节元件到转基因中或用大的基因组构件作为转基因而不用小的 cDNA 来避免。

精确地确定建立染色质结构域的元件很困难，但是有一些元件诸如边界元件、基质附着区和基因座控制区已被成功地应用于一些转基因（节 10.5）。为了更精确地研究这些长区域的作用和为了探讨人类基因在它们自己的顺式作用调节元件的背景下的基因表达和调节，建立允许大的 DNA 分子转移的环境是必要的。这方面主要的突破包括：

- **YAC 转基因小鼠**（YAC transgenic mice）的发展（Lamb and Gearhart，1995）。第一个发表的 YAC 转基因报告是关于一个 670kb 的 YAC 载体的转基因鼠，该载体含有人次黄嘌呤鸟嘌呤磷酸转移酶（*HPRT*）基因（Jakobovits *et al.*，1993）。这项技术对于建立由大规模剂量失衡引起的人类疾病模型和其他应用是非常有用的（节 20.4.5 部分），诸如，真正的小鼠的人抗体的制备（Mendez *et al.*，1997；关于一般方法，见图 20.6）。YAC 转基因可通过细胞融合、显微注射或脂质体转染等方法来产生。
- **转染色体小鼠**（transchromosomic mice）的发展（Tomizuka *et al.*，1997）。这些小鼠含有人染色体或染色体片段，是通过微细胞介导的基因转移制备的（框 8.4）。

20.2.4 基因转移也可用于产生特定突变和破坏内源基因的表达

迄今，我们仅从向动物细胞添加功能的观点来考虑基因转移技术。同样的技术也能用于其他方面，也许它对于生物医学研究的主要贡献是选择性删除或改变内源基因的功能。这不仅是研究基因功能有力的方法，也是制备精确模拟相应人类疾病模型的有用的

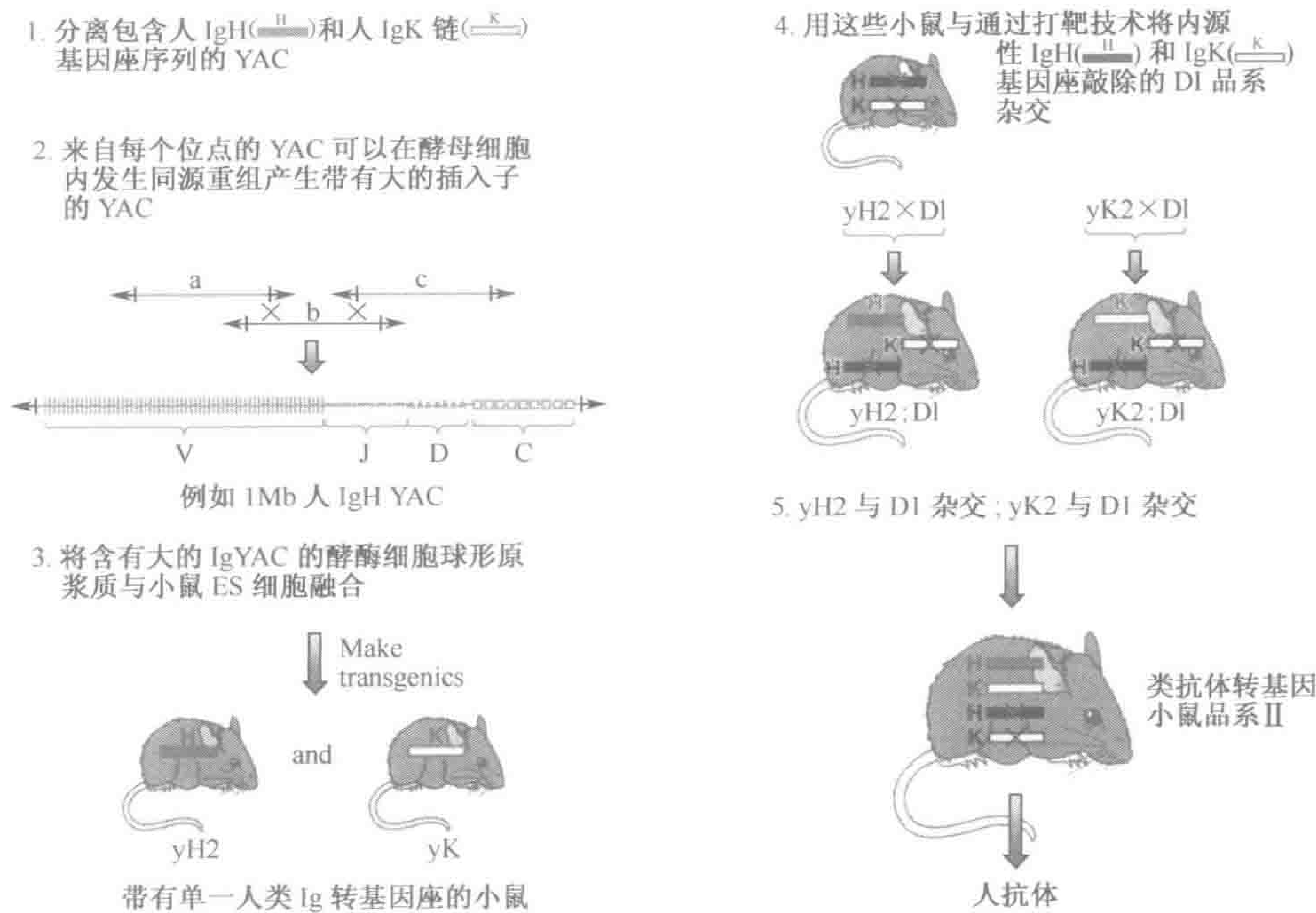


图 20.6　用 YAC 转基因构建带有人抗体的小鼠

YAC 含有免疫球蛋白（Ig）序列是通过用适合的 Ig 探针从 YAC 文库中筛查得到的。带有相对较小的插入子的 YAC 的恢复意味着有需要通过在酵母细胞中同源重组人工构建较大的 YAC 载体。通过处理酵母细胞建立球型原浆质，以致细胞外壁剥离掉，使细胞更容易融合。见 Memdez 等（1997）进一步叙述了制备这些小鼠的详细方法。

技术路线方法（节 20.4.4）。基因转移可通过三种方式修饰内源基因的功能。

- **基因打靶**（gene targeting）在这种方法中，用一个相关的序列应用同源重组以替代内源基因的序列，从而在基因组预定的位点导入一个特定的突变。这种方法通常只是用一个大的插入序列来简单的破坏基因，产生被称之为**基因敲除**（gene knock-out）的无效突变。然而，修饰的策略可用于更复杂的基因操作，包括精细突变的导入和一个基因替代另一个基因。
- **基因表达的抑制**（inhibition of expression）在这种方法中，标准的基因转移方法学是向细胞内添加一新的 DNA 序列。然而这个转基因产物的功能只是抑制内源基因的表达，而不是编码一个赋予新功能的蛋白。这些基因产物，包括反义 RNA、双链 RNA、小干涉 RNA、核酶、抗体和结构域的干扰蛋白，也可以直接导入而不是通过转基因来表达，虽然这种情况效果是瞬时而不是永久的。
- **插入诱变**（insertional mutagenesis）在这种方法中，转基因整合（随机的）到一个存在的基因并破坏它的功能。这种方法与基因打靶相似，通过导入一个突变来改变基因的功能，但这种情况下的突变既不是明确的，也不是靶向特定的位点。特定基因的突变可通过大规模的筛查来确定（节 20.3.3）。

20.2.5 基因打靶可产生每个细胞都带有特定突变的动物

基因打靶涉及内源基因和含有靶载体的外源 DNA 序列之间在体内的同源重组

导入动物细胞的 DNA 和宿主基因组之间的相互作用常导致外源 DNA 在预先存在的染色体缺口和断裂的位点发生随机整合，然而如果外源 DNA 序列与内源基因很相似，外源序列与内源序列之间会发生不同的相互作用，它们会发生排列结合并进行**同源重组**（homologous recombination）。这个过程，称作**基因打靶**（gene targeting），它可以在宿主基因组内预先选择的位点产生明确的突变。因此也被看作是体内人工定点诱变的一种形式（相对的各种体外定点诱变的方法在节 5.5.2，节 5.5.3 已作描述）。

同源重组在哺乳动物细胞内是很罕见的，约发生 10^4～10^5 次，比随机整合少很多。同源重组的频率依赖于同源区域的长度（打靶载体上与内源基因组合的部分）和与打靶基因的相似程度。因此，导入的经过修饰带有预期突变的 DNA 克隆一般是与宿主细胞的基因组 DNA 等基因的一段长的序列。即使这样，真正的同源重组事件的发生还是很少的，并且在相当大的随机整合背景下很难鉴别。用于细胞内基因打靶发生的策略依赖于构件的设计（图 20.7）。有两种类型的构件。

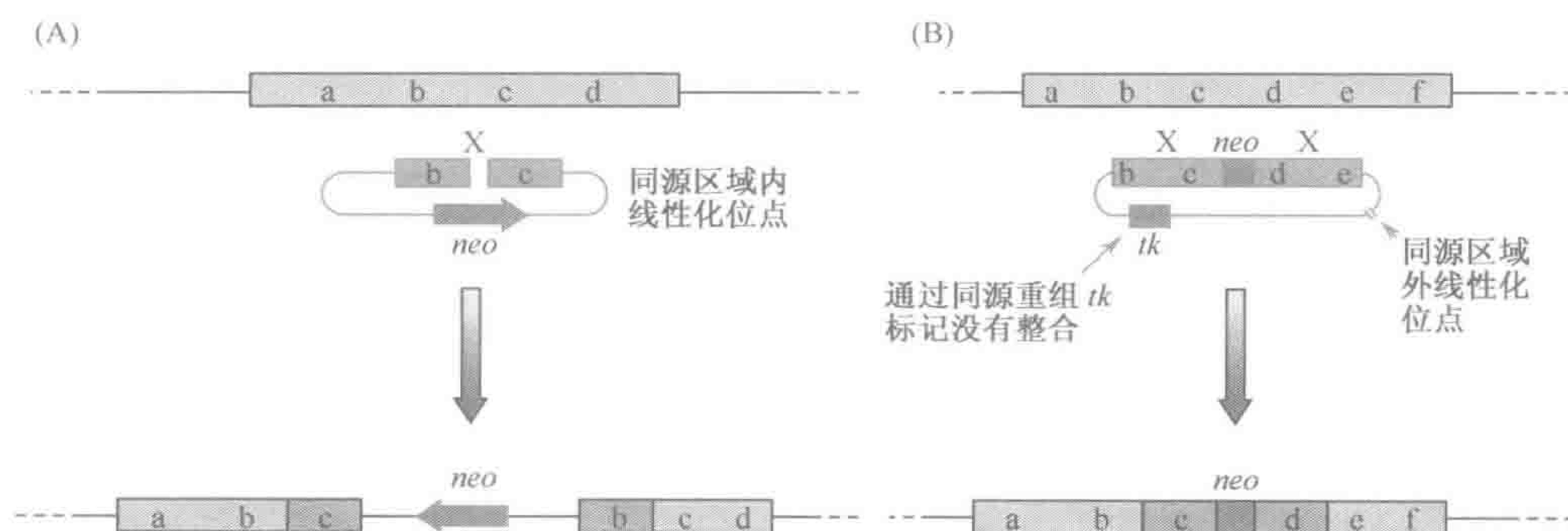


图 20.7 同源重组基因打靶可使完整的细胞内预定位点的基因失活

(A) **插入载体法**（insertion vector method）。导入的载体 DNA（红色）在单一位点切断，该位点的序列与打靶的内源基因（蓝色）部分密切相关或一致。通过同源重组（X），全部载体序列（包括标记基因 *neo*，对抗生素 G418 有耐药性）可插入到靶位点。然而真正的打靶事件很少发生，而在大部分抗 G418 的宿主细胞中载体发生随机整合。必须用 PCR 实验区别打靶和随机整合事件，在策略上可设计与载体和靶基因复性的引物。(B) **替代载体法**（replacement vector method）。这种情况，*neo* 基因包含在与内源基因同源的序列内，载体在同源区域外的单一位点切开。通过与载体同源序列发生两次重组或基因反转事件（XX）导致靶基因内部序列的替代，包括 *neo*。另外，随机整合事件更常见，但替代载体法适用于更复杂的两步筛选策略，其中的第二个标记是一负性筛选标记，诸如 *tk*（对 *ganciclovir* 敏感）在同源区域外侧，可通过随机整合而非同源重组整合到基因组中。因此，同时抗 G418 和 *ganciclovir* 的细胞可能是正确的靶标。字母表示基因的线性顺序不代表外显子。

- **插入型载体**（insertion vector） 通过单一相互重组事件打靶感兴趣的位点，发生整个载体的插入（图 20.7A）。
- **置换型载体**（replacement vector） 设计导入 DNA 的一段同源序列置换染色体上的一些序列（图 20.7B），这可通过两次相互重组或基因转换而实现。

在这两个策略中，一个含有新霉素（*neo*）标记的大片段 DNA 载体导入打靶位点后可引起基因紊乱并产生无效等位基因（**基因敲除**，gene knockout），然而这不可能总是令人很满意，如果需要更精细的突变，可用各种两步重组技术。如图 20.8 所示的用插入载体的“打了就跑”策略和用替代载体的“标记和置换”策略。

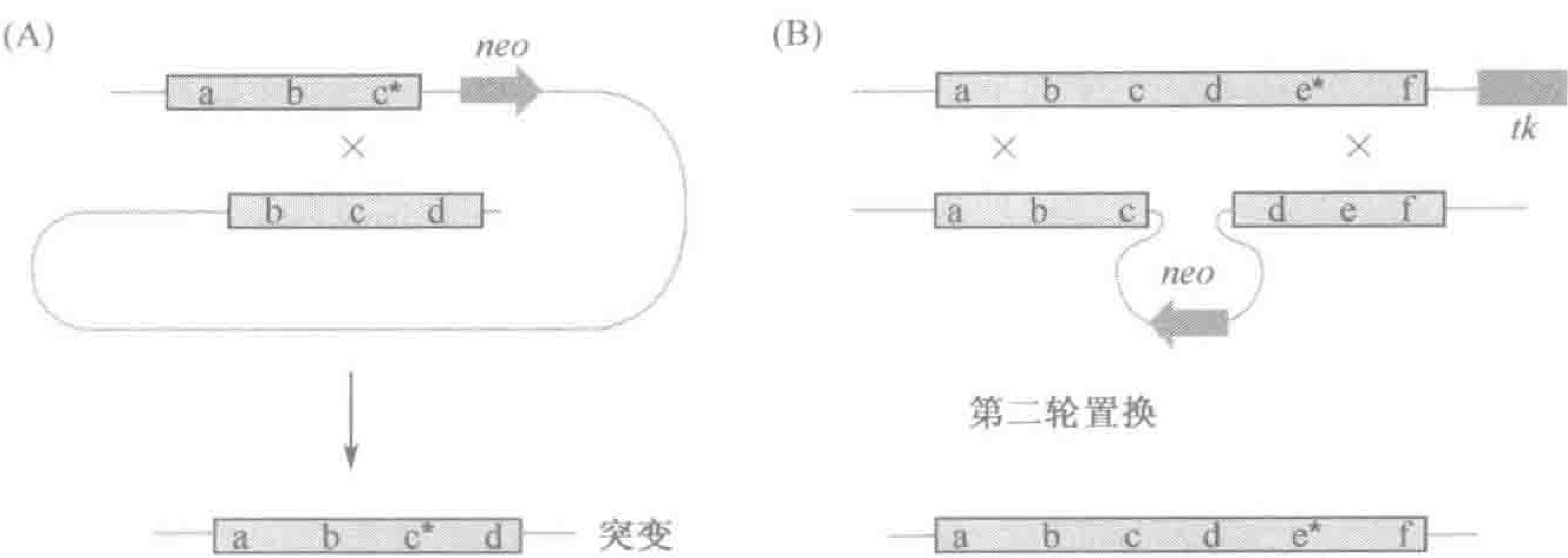


图 20.8　通过基因打靶导入精细突变

(A) 在应用插入载体的“打了就跑”策略中，精细突变存在于第一次打靶的构件上，染色体内重组导致标记基因和载体骨架的消除。(B) 在应用替代载体的“标记和置换”策略中，第二次打靶构件用于替代第一次导入的突变，第二次的构件在同源区域外带有一个负性筛选标记以避免随机整合。

遗传修饰的动物可用基因打靶 ES 细胞或用体细胞做供体进行核移植的方法来制备

通过同源重组进行的基因打靶首先从培养的人/小鼠杂种的体细胞（Smithies *et al*., 1985）获得，并且此后在许多哺乳动物细胞中得到证实。然而，基因打靶最重要的是小鼠 ES 细胞的应用，ES 细胞对打靶技术具有特殊的作用，因为在 ES 细胞同源重组发生率相对较高（虽然仍比随机整合率低）。ES 细胞不同寻常又极为有益的三个特点的结合——容易培养和转染，易发生同源重组，具有多能性——意味着直接用 ES 细胞可以在每一个细胞里都产生带有相同遗传修饰的转基因鼠（Capecchi，1989；Melton，1994）。只要打靶 ES 细胞制备好就可以像图 20.3 所示那样按常规的方法制备转基因小鼠。这样的小鼠称为“遗传修饰”或“打靶”小鼠而不是转基因鼠，因为它们不一定包含外源 DNA。确实，在预先确定的位点产生包含单个点突变的打靶小鼠是可能的。

尽管体细胞基因打靶方法的效率普遍很低，但最近靶向修饰的成纤维细胞已经用作核移植的供体（节 20.2.2）。这导致了许多遗传修饰的哺乳动物的产生。最初的报告描述了在 *COL1A1* 基因座导入一个新基因的山羊（McCreath *et al*.，2000），最近，有两个小组报道了靶向破坏 α-1，3 半乳糖转移酶基因猪的产生（Dai *et al*.，2002；Lai *et al*.，2002）。这种酶可用羟基修饰蛋白，该酶在灵长类没有发现，是猪-人器官移植时排斥反应的一个主要反应因子。

20.2.6　特定位点重组可用于条件基因失活和染色体工程

特定位点重组（site-specific recombination）系统见于一些噬菌体、细菌和酵母。每个系统包括两个小的部分：一个是在其中发生重组的短的特异识别序列，一个是重组酶，它识别这一序列并在它的两个拷贝存在时执行重组反应。特定位点重组能力是在于

识别位点能够容易通过基因工程手段导入到转基因或靶载体中，而重组酶可有条件地提供，因为编码重组酶的基因可在调节或诱导型启动子的控制下表达。因此至今，来自P1 噬菌体 *Lox*P 的 Cre-*lox*P 重组系统已被广泛地应用，尤其是在遗传修饰的小鼠中。**Cre 重组酶**（causes recombination）（引起重组）的天然功能是介导两个 34bp 的之间的重组，*Lox*P 序列是由中间的非对称的 8 个碱基间隔的两个 13bp 的反向重复序列组成（图 20.9）。如果两个 *lox*P 位点方向相同，它们之间的间隔序列被删除，如果方向相反，它们之间的序列就发生反转，因此 Cre-*lox*p 系统被应用于一些不同的方法中，包括转基因的特定位点整合、转基因的条件激活和失活以及不想保留的标记基因的删除等。然而，最重要的应用也许是下面讨论的条件基因失活和染色体工程（Lobe and Nagy，1998）。



图 20.9 *Lox*P 识别序列的结构
注意这个两侧带有 13 个碱基的反向重复序列的中间 8 个碱基序列是不对称的，并有方向性。

条件基因失活

一些基因在发育早期起关键作用，简单的敲除实验一般是无用的，因为在胚胎早期阶段就确有死亡。为了克服这个问题，开发了只在动物特定的细胞或其发育的特殊阶段使靶基因失活的方法，因此这样动物可以生存，并且可以对感兴趣的组织或细胞类型研究**条件敲除突变**（conditional knockout mutation）的效应。这种方法一个早期的例子是由 Gu 等（1994）报道的 DNA 聚合酶 β 的条件敲除。这是胚胎发育一个必需的酶。基因打靶方法是用一个两侧带有 *lox*P 序列的同源基因片段替代内源基因的一个重要的外显子。带有这种靶突变的小鼠与带有在 T 淋巴细胞特异性启动子控制下的 Cre 转基因品系小鼠交配。杂交的后代被确认含有两个转基因并且生长到成年。Cre 产物只在 T 淋巴细胞表达，删除靶基因的重要外显子并使其失活。（见图 20.10 方法）。这种方法的一般优点是 Cre 转基因鼠可以被反复用于不同的实验。无论 *lox*P 位点中间是什么内源基因，都可用上述的 Cre 转基因小鼠来破坏 T 细胞里的这个基因。同样，广泛应用的 Cre 转基因已经用于四环素诱导表达和与雌激素受体融合的组成性表达系统中，并在蛋白水平可被枸橼酸他莫昔芬（Tamoxifen）激活（节 20.2.3；Fiel *et al.*，1996）。

染色体工程

另一个近来重要的进展是在 ES 细胞中依靠连续基因打靶和 Cre-*lox*P 重组的染色体工程策略，基因打靶用于在染色体的预定位点整合 *lox*P，然后 Cre 重组酶瞬时表达，介导一个选择性染色体重排（Ramirez-Solis *et al.*，1995；Smith *et al.*，1995；图 20.11）。这种染色体工程策略为建立用于遗传研究的带有特定染色体畸变的新小鼠品系提供了令人兴奋的可能性。采用 Herault 等（1998）的新方法可避免上述染色体工程中 ES 细胞的多次打靶和筛选过程。这个靶向减数分裂重组方法利用了自然情况下发生在

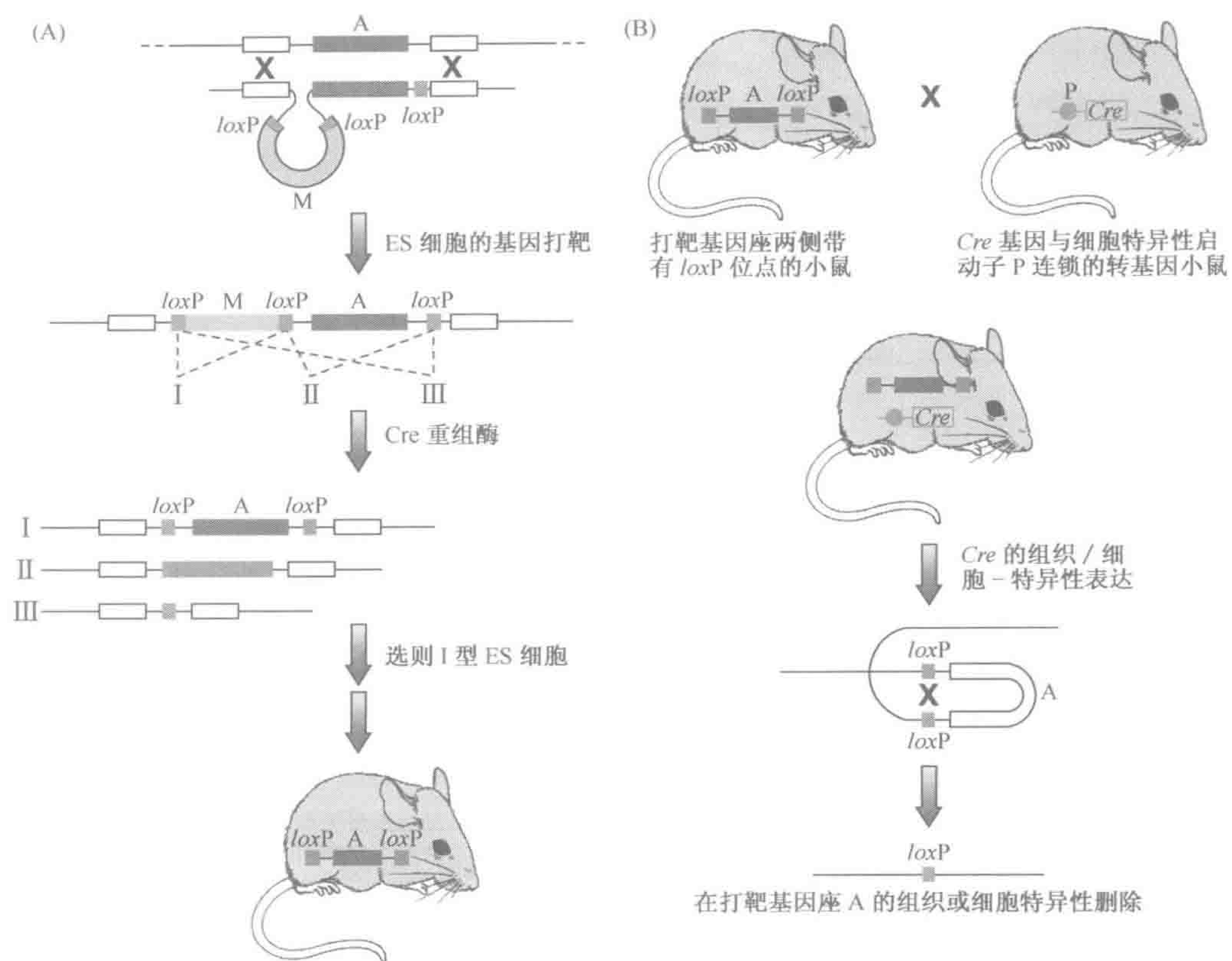


图 20.10　Cre-*loxP* 重组系统基因打靶可用于特定类型细胞的基因失活

(A) 用小鼠 ES 细胞阐述一个标准同源重组方法，三个 *loxP* 位点和一个 M 标记基因一起导入靶基因座 A（一个典型的小基因或一个固有的外显子，如果它缺失可引起移码突变）。随着 Cre 重组酶基因的转染和瞬时表达导致 *loxP* 位点间的重组，产生不同的产物。1 型重组用于产生两侧是 *loxP* 位点的靶基因座小鼠，这样的小鼠可以与以前构建的转基因小鼠进行交配。(B) 这个小鼠携有由 Cre 重组酶基因与特定组织的启动子组成的整合构件。使含有两个 *loxP* 侧翼序列靶基因座和 Cre 基因的后代在所期望类型的细胞内表达 Cre 基因，这些细胞中 *loxP* 位点间的重组导致靶基因座 A 在特定组织的失活。

细胞第一次减数分裂中同源染色体配对的特点。设计一个转基因在 *Sycp1* 启动子（*Sycp1* 基因编码部分联会复合体，有利于交叉的发生）的控制下表达重组酶。结果，Cre 重组酶在雄性精原细胞从偶线期到粗线期阶段染色体配对时产生。

20.2.7　转基因策略可用于抑制内源基因功能

尽管基因打靶毫无疑问是在细胞和转基因动物中进行基因功能操作的最精确和最直接的方法，但一个主要的局限性是此方法只可应用于对小鼠和果蝇的常规基础研究。因此，开发了更适合普遍应用的抑制基因的其他方法（表 20.2）。尽管这些方法是不同的，但把它们放在一起考虑，因为它们都是以某种方式干涉基因表达或功能，而没有改变这个靶基因的 DNA 序列。有的时候它们被称作**功能敲除**（functional knockout）方法，因为它们产生了相应的突变表型的**拟表型**（phenocopy）。

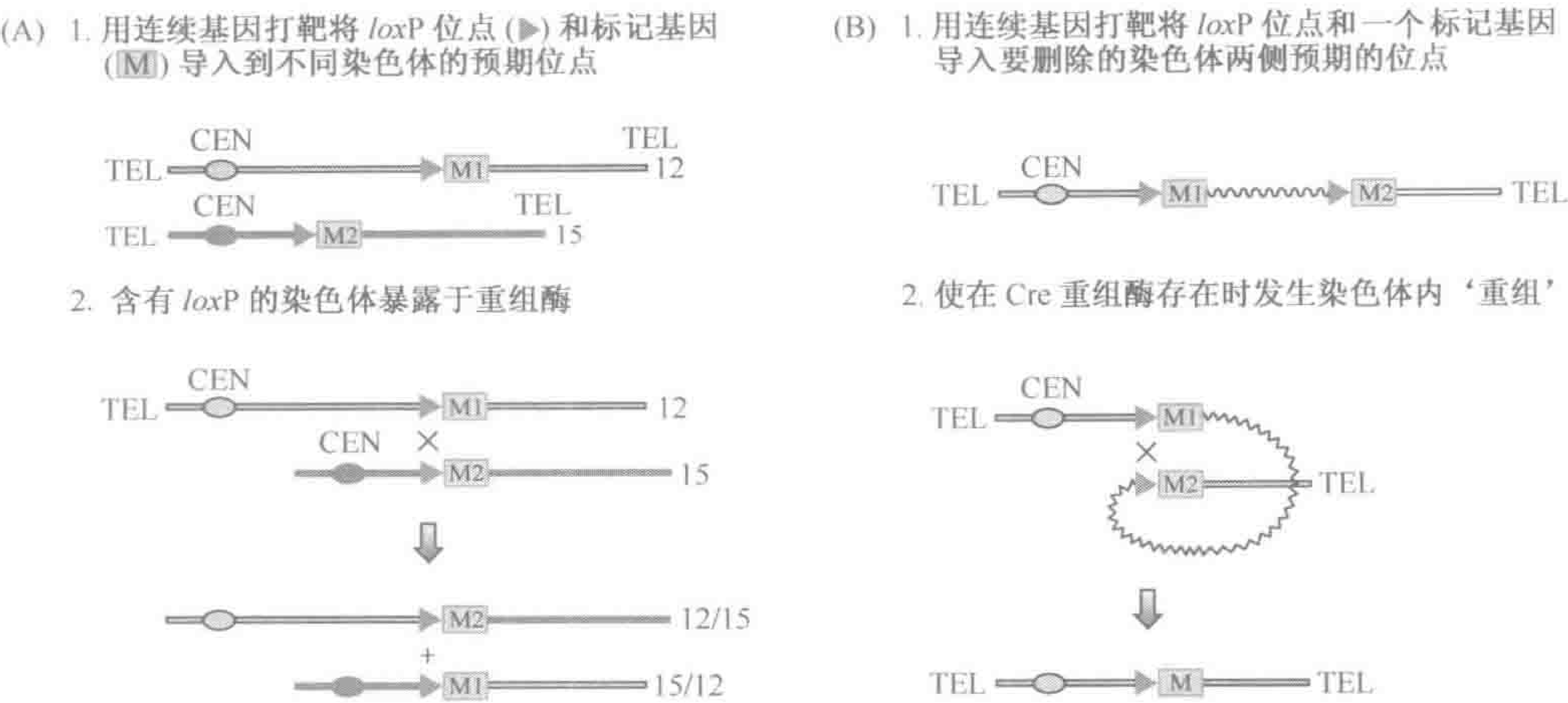


图 20.11 用 Cre-*loxP* 系统可以完成染色体工程

(A) 应用靶向插入 *LoxP* 位点，促进染色体易位，见 Smith 等（1995）的一个实例。(B) 应用靶向插入 *LoxP* 位点，可制作染色体内重组的微缺失。见 Ramirez-solis 等（1995）的一个实例和其他染色体内重组的例子。

表 20.2 没有突变的靶基因干涉内源基因表达方法的概述

RNA 水平干涉	蛋白水平干涉
反义 RNA	显性负性改变
反义寡核苷酸	抗体，细胞内抗体
核酶或大核酶	适体，细胞内适体
脱氧核酶	
正义 RNA(共抑制)	
dsRNA(RNA 干涉)	
siRNA((RNA 干涉)	

基因表达可通过靶向破坏特异性 mRNA 来抑制

RNA 是基因和蛋白质之间的桥梁，因此可通过选择性破坏或失活特异基因的转录物以产生功能性敲除。有几种方式可使 mRNA 靶向特异性失活，所有这些方法最终都涉及至少部分双链 RNA 分子的形成。有些特定的基因受内源性的反义 RNA 称作**小瞬时 RNA**（small temporal RNA）的调节（stRNA，节 9.2.3）。它们看上去是通过与转录物的结合来抑制 mRNA 的加工或蛋白质的合成。因此早期基因抑制策略是向细胞导入**反义 RNA**（antisense RNA）或**反义寡核苷酸**（antisense oligonucleotide）。这些分子直接导入后可见瞬时的抑制效应，而永久的抑制效应要通过用带有能持续产生反义 RNA 的**反义转基因**（antisense transgene）的转化细胞或动物来实现。这个方法最早是被 Katsuki 等（1988）在小鼠上证实的，他成功地通过反义 cDNA 使磷脂蛋白表达减少到正常水平的 20%，产生了颤抖突变体表型。也可以用诱导型启动子表达反义 RNA 来调节培养细胞的生长（Sklar *et al.*，1991）。

原来认为反义 RNA 的效果是计量的（即一个反义分子阻止一个转录本的翻译，然

后两者一起被降解)。因此，如果抑制分子被循环利用会有更大的抑制功效，以至于抑制分子自身被降解之前许多转录物被破坏。**核酶**（ribozyme)，一种催化剪切 RNA 分子的酶被发现有这个特点（节 21.6)。已经设计了含有核酶催化中心整合在反义转基因的构件，以促进特定转录物的破坏。这样的构件已被广泛应用于细胞系，尤其用于癌基因的抑制和抵制 HIV 感染的研究（Welch *et al.*，1998)。但它们在转基因鼠上少见表达。一个有用的例证是在胰腺的 β 细胞内通过控制胰岛素启动子的反义葡萄糖激酶核糖酶转基因来特异性靶向抑制葡萄糖激酶 mRNA 来建立的糖尿病模型（Efrat *et al.*，1994)。最近开发了通过构象调节控制核酶的活性（称为**"大核酶"**，maxizyme）并已被用于有条件地抑制基因表达（Kuwabrara *et al.*，2000；Famulok and Verma，2002)。

令人吃惊的是核酶构件在大多情况并没有显示出比相应缺乏核酶催化中心的反义 RNA 更好的抑制效应，这表明反义 RNA 可有比想象的单独计量结合更强的作用。这一现象的基本线索是在某些情况正义 RNA 在哺乳动物细胞的表达具有抑制相应内源基因的表达的能力，这种现象被称作**共抑制**（cosuppression)（Bahramian and Zabl，1999)。共抑制已经在植物中被广泛证明，并涉及异常 RNA 种类，尤其双链 RNA 的形成。对线虫（*C. elegans*）中正义 RNA 和反义 RNA 沉默基因能力的研究发现了一种称为 **RNA 干涉**（RNA interference）的新现象，即对某一特殊基因同时导入正义和相应的反义 RNA 会导致高效、持久和非常特异的基因沉默（Fire *et al.*，1998)。RNA 干涉是高度保守的细胞防御机制也发生在哺乳动物细胞（包括人类)。它是由**双链 RNA**（double-stranded RNA，dsRNA）触发并引起与诱导分子序列相同的单链 mRNA 降解。

RNA 干涉的机制是复杂的，它涉及以下一个过程：一个 dsRNA 特异性的内切核酸酶将 dsRNA 分子降解为短的长约 21～25bp 的双链 RNA 分子（图 20.12)。这短的双链称为**小干涉 RNA**（small interfering RNA，siRNA)。这些分子与相应的 mRNA 结合并装配一个序列-特异性 RNA 内切核酸酶称为 **RNA 诱导的沉默复合体**（RNA induced silencing complex，RISC)，它能非常有效的将大多数基因的 mRNA 减少到难以检测的水平。有趣的是，已知同样的 Dicer 酶也加工上面讨论的小瞬时 RNA 从而干扰内源基因的表达。在许多其他的生物体中，包括人类，也发现了类似的具有内源性调节功能的 RNA 分子，叫做微 RNA [（miro-RNA，miRNA)，节 9.2.3]。一般认为 miRNA（阻止蛋白翻译）和小干涉 RNA（催化降解 mRNA）功能的不同反映它们前体的结构和在 Dicer 处理之前其他酶的结构的不同。Micro RNA 是单链的，来自带有突出和环状的不完全双链结构 RNA，而 siRNA 是双链的，并且通常是来自完全双链的 RNA（Pasquinelli，2002；Voinet，2002)。这两个途径之间可能有交叉的地方（图 20.12)。

RNA 干涉机制可以用于细胞也可以用于胚胎，因为它是一个全身的现象——siRNA似是能在细胞之间移动，因此 dsRNA 导入部分胚胎就可引起全部彻底的沉默。像 dsRNA 直接导入细胞（通过转染）或胚胎（通过注射或其他方法）一样，也可以表达导入两个正、反义链 RNA 的转基因或导入一个产生反向重复序列的结构的构建，该构建可表达作为 Dicer 的底物的发夹样 RNA。

在大部分动物和哺乳动物胚胎和胚胎细胞系导入或表达长的 dsRNA 分子适合于基因沉默。然而，在成年哺乳动物细胞，RNA 干涉结果被**干扰素反应**（interferon response）所掩饰，这是对 30bp 以上的双链 RNA 产生的一般的（非序列特异性）反应。

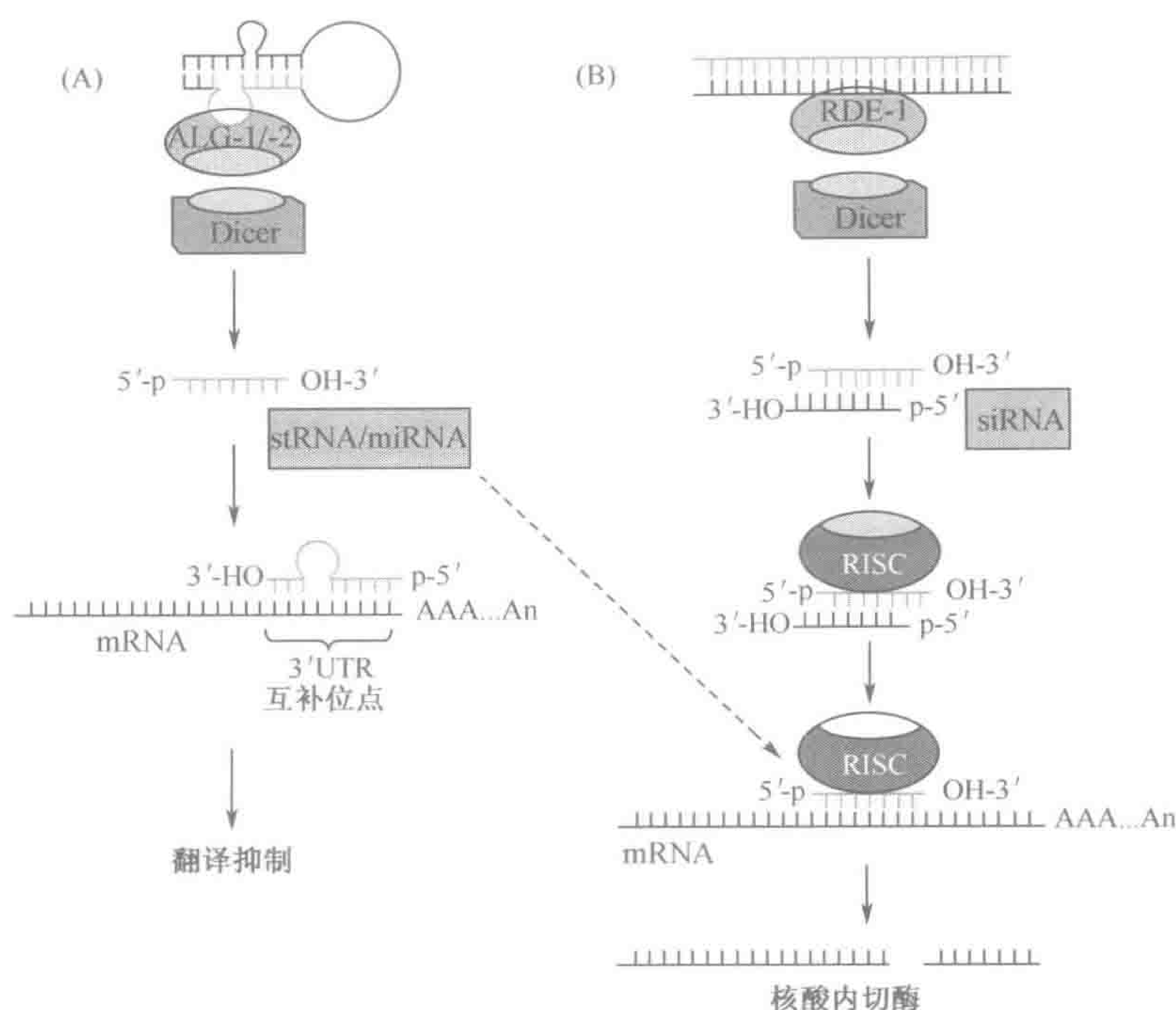


图 20.12 在秀丽新小杆线虫中 stRNA（miRNA）和 siRNA 途径的比较

通过识别特异类型的前体 RNA，相关蛋白质诸如 ALG-1/ALG-2 和 RDE-1 可通过 DICER 部分地影响剪切过程。（A）stRNA（miRNA）是单链结构，与靶 mRNA 的 3′UTR 不完全配对抑制翻译的启始。（B）siRNA是双链且 3′有两个核苷酸突出末端，并整合到 RNA 诱导的沉默复合体（RISC）。此 siRNA 指导 RISC 与 mRNA 精确配对，靶 mRNA 在与 siRNA 形成双链的中间部位被剪切。miRNA 有可能与它们的靶序列精确配对被导入 RISC。修改自 Voinet（2002），经 Elsevier Science 同意。

这个问题已经通过直接用化学的或酶促合成的 siRNA 或转小基因表达法得到解决，一般根据设计小 RNA（例如 U6snRNA 基因）的内源基因。经过最近 3 年时间，对 RNA 干涉的认识有了提高，从相对模糊到可能是在细胞和动物体内进行高通量功能分析的最有前途的工具（见下文），并且目前可发展成为整个新一类治疗药物的基础（节 21.6）。见 Tuschl 和 Borkhardt（2002），这是一篇关于 RNA 干涉和应用的综述。

基因功能也可以在蛋白水平阻断

即使一个转录物有活性并被翻译，基因功能也可以在蛋白水平被阻断，产生功能缺失模拟表型。如果靶向内源基因的产物发挥多聚体作用，就有可能在细胞或转基因动物产生一个高水平表达的**显性负效突变体**（dominant negative mutant）。这种情况，蛋白的功能拷贝被隔离以失活复合体的形式存在。这种方法已被广泛应用于阻断受体功能，因为许多受体是以二聚体发挥功能（例如 Amaya *et al.*，1991）。

此外，可以表达识别靶蛋白的抗体中和它。功能性失活可通过直接向细胞或动物体内导入抗体来实现，或由一个转基因表达抗体，这种情况通常称作**细胞内抗体**（intrabody）（Richardson and Marasco，1995）。DNA 或 RNA 寡核苷酸也能与蛋白结合，并

以某种特殊形式阻抑蛋白质活性。这些称为**适体**（aptamer），如果是在细胞内表达，则称为**细胞内适体**（intramer）（Famulok *et al*.，2001）。

20.3　利用基因转移研究基因表达和功能

20.3.1　可用报道基因研究基因表达和调节

转化的细胞系和转基因生物最早的用途之一是**基因调节**（gene regulation）的分析，即鉴定基因表达所需的特异性序列。通过研究基因上游或偶尔在第一内含子的不同 DNA 片段的缺失如何影响其表达可以研究基因调节。显然，人类的细胞是研究人类基因表达最合适的系统，在导入含有不同量的 DNA 侧翼序列的构件，然后进行基因表达的分析是一个合乎逻辑的方法。然而，这有一个问题是在内源性同源基因存在并可能在同一细胞内表达的情况下导入的人类基因如何表达。要解决这个问题，将推测的调节序列克隆到**报道基因**（reporter gene）的上游载体上，这个基因产生的蛋白就可通过一个简单的方法和定量进行检测（框 20.4）。

框 20.4　动物细胞的报道基因

报道基因编码的蛋白质可以用简单的低廉的实验进行检测并定量。它们可用于观察和测量基因转移及表达的效率，并可研究细胞内蛋白质的定位。几个报道基因常用于动物。

来自 *E. coli* 转座子 *Tn9* 的 *cat* 基因编码**氯霉素乙酰基转移酶**（chloramphenicol acetyltransferase），它能从乙酰辅酶 A 将乙酰基团转移到氯霉素上。现在已有标准的体外实验方法（Gorman *et al*.，1982）。细胞裂解与标记^{14}C 的氯霉素混合、孵育，然后通过薄层层析法分离。用磷酸成像仪或闪烁仪检测乙酰化和非乙酰化氯霉素的相对量来决定 CAT 的活性。

来自于 *E. coli* 的 *lacZ* 基因编码的 **β-半乳糖苷酶**（β-galactosidase）能够降解乳糖和相关的成分。若干特异性衍生底物包括 ONPG 是可用的，ONPG 产生可溶性黄色产物并用于体外实验的比色，而 X-gal 产生蓝色沉淀并用于原位检测（Hall *et al*.，1983）。这个基因大多同 X-gal 一起用做组织标记显示转基因动物的基因表达模式。*E. coliin gusA* 基因编码 **β 葡萄糖苷酸酶**（β-glucuronidase），使用方法与 *lacZ* 相似。

来自于美国萤火虫（*Photinus pyralis*）的 *Luc* 基因编码**萤光素酶**（luciferase），它在有氧、ATP 和镁离子的条件下催化荧光素氧化（de Wet *et al*.，1987）。该反应释放闪光，光的强度与酶的活性水平成比例。可以用荧光仪和闪烁计数仪进行检测，该实验比传统的 CAT、β-半乳糖苷酶或 β-葡萄糖酸酶敏感 100 多倍。光释放信号减弱很快，因此萤光素酶用于监测基因表达水平的快速改变。相反，CAT、β-半乳糖苷酶或 β-葡萄糖酸酶是很稳定的蛋白质，所以当基因打开的时候可以有效地进行检测，但基因关闭后仍然持续。来自于其他生物的萤光素酶基因活性略有不同并且反应释放的光波长也不同，可用于同时监测几个基因。

Gfp 基因来自于维多利亚发光水母鱼，编码**绿色荧光蛋白**（green fluorescent protein，GFP），是一种生物荧光标记，当暴露于蓝色或紫外光下时释放明亮的绿色荧光（IKawa *et al*.，1999）。GFP 的活性可通过对释放的绿光强度来定量检测，但与萤光素酶不同，该蛋白质不需要底物，可以很容易的对细胞的过程进行实时监测。GFP 融合蛋白广泛用于细胞内蛋白定位和细胞内或细胞间蛋白质的运输。许多不同的 GFP 可用于双标记。其他颜色的荧光蛋白质也是有用的。一种由绿色荧光蛋白突变的红色荧光蛋白可超时用于瞬时基因表达的分析（Terskikh *et al*.，2000）。

图 20.13 所示在细胞内用定位调节元件的常规方法。使用不同的细胞系，一些细胞系内表达相应的内源基因，而有些则不表达，这可能提供了在许可的细胞系内允许基因表达的启动子元件和在不许可的细胞系内阻止基因表达的启动子元件的证据。同样，对可诱导基因表达的元件可通过在同一细胞系里在有无诱导剂的条件下检测不同的启动子构件进行定位。细胞特异性基因表达的一个更精确的分析可以通过研究转基因动物体内报道子的表达模式，因为这种方法可以鉴定在特殊细胞类型或发育的不同阶段控制基因表达的元件。然而必须通过来自几个独立的转基因品系得到相似的结果才能避免由位置效应导致的对表达模式的错误解释（节 20.2.3）。

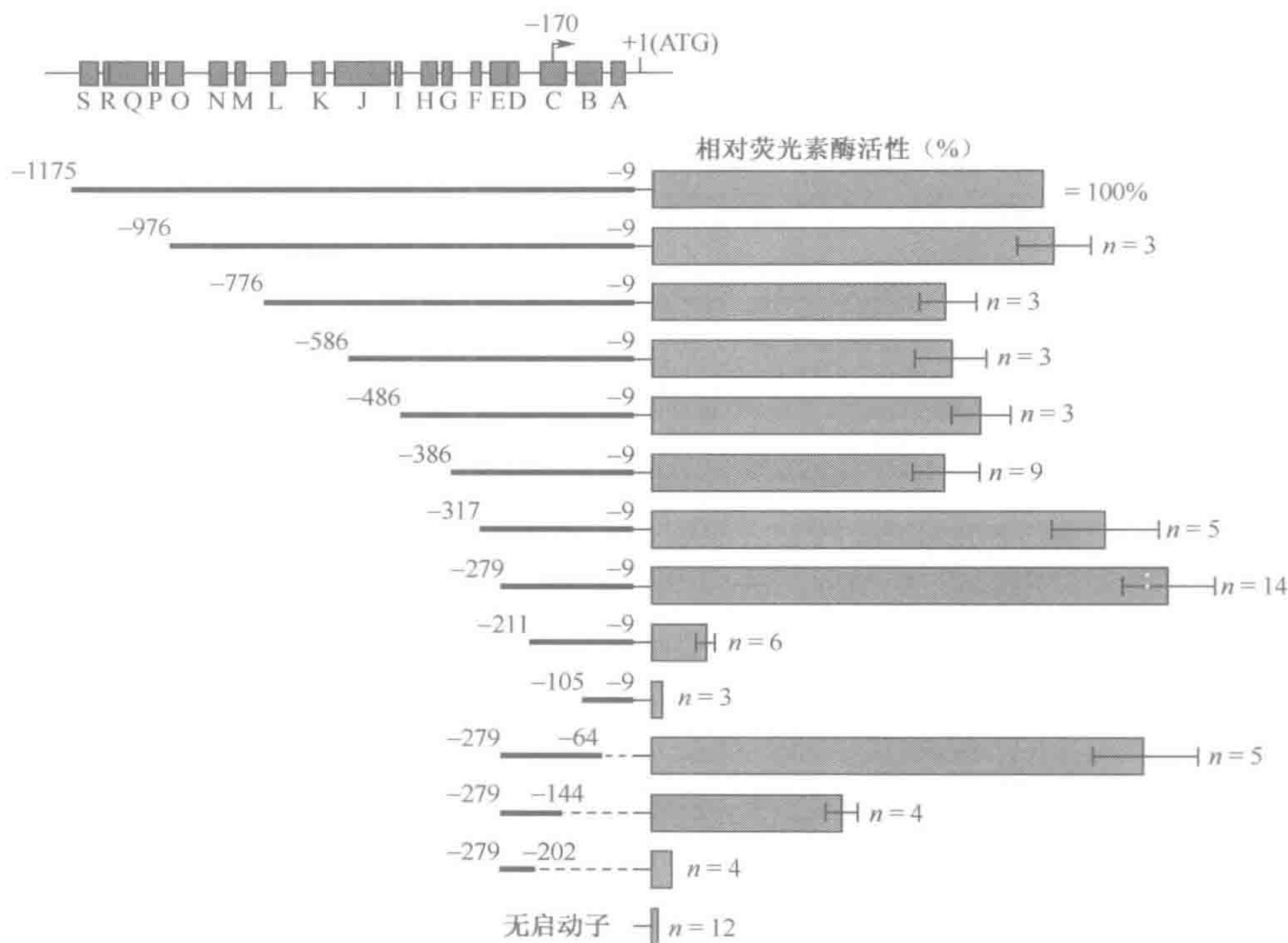


图 20.13 人Ⅷ因子启动子区的缺失分析

如果人的基因表达所需要的必要元件都存在，当构件转染到适当类型的人细胞中，可导致报道基因高水平表达。用限制性内切酶或来自 *E. coli* 外切核酸酶 Ⅲ，从双链 DNA 3′端消化而制备一系列缺失构件，或者消化一系列 PCR 扩增产物，使其包含感兴趣区域，然后克隆到合适的表达载体中，左侧实棒表示人Ⅷ因子（F8）基因上游不同大小的序列。最上面的框表示蛋白结合位点的上游序列。右面的框表示基于 *n* 重复实验相关的完整序列的萤光素酶活性水平。根据不同表达克隆观察到的萤光素酶活性，序列缺失图谱显示维持启动子最大活性的所有必需的元件位于－279～－64 区域，包括蛋白结合位点 B 、C 和 D。重绘源自 Figueiredo 和 Brownlee (1995)，经 American Society for Biochemistry and Molecular Bioligy 同意。

20.3.2 基因功能可通过产生功能丢失和功能获得突变和拟表型来研究

由于遗传的丰余基因功能不总是可以通过破坏或抑制来建立

在细胞水平研究可能的情况下，RNA 干涉成为选择用来产生功能丧失效应的方法

（节 20.2.6）。许多基因包括已表明是 HIV 出芽必须蛋白的编码血管支架蛋白 Tsg101 基因的功能都已通过对培养的哺乳动物细胞进行 RNA 干涉来确立（Garrus *et al*.，2001）。Tuschl 和 Borkhardt（2002）已引证了很多其他的例子，包括细胞分化、DNA 甲基化、细胞信号和膜转运等基因的分析。在完整的生物背景下从事基因功能研究最直接的方式是基因打靶（节 20.2.4）。许多情况下，这些突变提供非常多的信息并揭示大量的基因功能。因此开展很多基因敲除实验并且所有的结果根据其表型分类在国际互联网的数据库上建立了目录（表 20.3）。然而在许多情况下，敲除突变显示极少的表型效果，这可能是由于**遗传丰余**（genetic redundancy）的原因，即还有另外的基因可以代替失活基因的功能。因此，为确定一些基因精确的功能需要敲除两个甚至三个基因。

表 20.3　哺乳动物转基因和诱发突变及大规模基因捕获和 RNAi 扫描的互联网资源

资源	URL
转基因小鼠和靶向诱发突变资源	
Jackson 实验室，打靶突变小鼠数据库	http://jaxmice.jax.org/index.shtm
TBASE，转基因和打靶突变小鼠数据库	http://tbase.jax.org
由 Jackson 实验室维护	
科学基因敲除数据库尖端	http://www.bioscience.org/knockout/knockhome.htm
生物医学网小鼠敲除数据库	http://biomednet.com/db/mkmd
小鼠 ENU 诱发突变资源	
德国人类基因组计划 ENU 诱发突变	http://www.gsf.de/ieg/groups/enu-mouse.html
MRC 碱基突变 ENU 诱发突变数据库	http://www.mgu.har.mrc.ac.uk/mutabase
基因捕获资源（小鼠和果蝇）	
德国基因捕获协作组	http://tikus.gsf.de
Lexicon 遗传学	http://www.lexgen.com/omnibank/omnibank.htm
Berkeleg 果蝇基因组计划	http://www.fruitfly.org/p disrupt/index.html
秀丽新小杆线虫 RNA 干涉资源	
一般信息	http://www.wormbase.org
RNAi	http://www.rnai.org

一个提供信息的例子是 *Myo D* 和 *Myf*-5 基因，它们在小鼠发育的早期表达。相应于这两个基因克隆的 cDNA 在许多不同的细胞系具有诱导肌组织-特异蛋白表达的能力。因为这两个基因都编码转录因子，他们看上去都是肌原蛋白调节子最好的候选基因。然而，奇怪的是当产生敲除突变时没有一种情况可以产生明显表型，肌肉发育在 *MyoD* 敲除的小鼠显然正常，而在 *Myf*-5 敲除的小鼠稍有迟缓（Rudnicki *et al*.，1992）。后来，发现这两个基因具有相同的功能，每一个基因可以完全补偿另外一个基因的缺如。的确，在正常的小鼠上，这两个转录因子相互彼此抑制，以至于一个基因的敲除可导致另外一个基因补偿性表达增强。*Myf-5* 敲除小鼠的轻微表型反应了它的表达要稍早些。两个基因都敲除的小鼠具有严重的肌肉发育障碍，该突变小鼠出生后由于

呼吸窒息而死亡。

如果基因产物的剂量、活性和分布至关重要可通过功能获得实验来确立基因功能

如果功能丢失突变或表型模拟不提供信息，作用于相同基因功能显示的转化细胞和转基因动物可能是有用的。在这一点上 *Myo D* 和 *Myf*-5 基因就是一个实例。由于它们对培养细胞的影响，已知这些基因编码重要的肌肉发育调节因子。同样，许多原癌基因根据它们引起培养细胞不可控制地增殖的特性已被鉴定并确定了功能。功能获得可通过一个正常基因产物的过表达或使该基因的过度激活突变体表达来建立。在转基因动物，功能获得实验可同样有用。经典的例子是 *Sry* 基因，通过其在含有两个 X 染色体的转基因小鼠中的表达显示该基因是主要的雄性发育决定子。这些小鼠原来是为雌性做准备的，反过来成了雄性鼠（Koopman *et al*.，1991）。另外一种功能获得是**异位表达**（ectopic expression），转基因在相应的内源基因正常时空领域之外进行表达。有许多用这种方法研究发育基因的有用的例子。例如 *Hox* 基因在其正常结构域之外表达破坏胚胎的正常模式导致肢体和骨骼的缺陷（Burke，2000）。

异位表达实验也可用来证明遗传丰余，例如，小鼠的齿状基因（engrailed）*En-1* 和 *En-2* 是含有同源盒基因，被认为是在脑的形成中具有关键作用。*En*-1 基因敲除突变体显示了和设想结果一致的严重的脑发育异常而 *En-2* 基因敲除小鼠只有轻微缺陷。*En*-1 表达比 *En*-2 表达早 8～10h，提示在 *En-2* 敲除小鼠可能是 En-1 蛋白表达补偿了 En-2 蛋白。为了检测功能丰余的可能性，Hanks 等（1995）使用了被称作**基因敲入**（gene knock-in）的一系列敲除程序，在这个程序中打靶构件含有在 *En*-1 同源区域（包括正常的 *En*-1 基因调节元件）内的 *En*-2 基因。其目的是用 *En-2* 的编码序列替代内源性 *En-1* 的编码序列，产生有 2 个 *En*-2 基因的小鼠，其中一个和正常胚胎中 *En*-1 的表达模式一样表达（图 20.14）。产生的 *En*-1 敲除的小鼠具有正常表型，证明敲入的 *En*-2 基因与 *En*-1 具有相同的功能。

20.3.3　应用插入诱变和系统 RNA 干涉大规模分析基因功能是功能基因组学研究的基础

上述实验是设计研究单个基因表达和功能的，然而，如 19 章讨论的，基因组计划中获得的大量序列信息和未知功能基因现在需要在全基因组范围研究基因功能。饱和诱变已用于功能研究很多年了，这些研究依赖于放射线或强的化学诱变剂如乙基亚硝基脲（ENU）和乙基甲硫酸盐（EMS）产生代表基因组内每个基因的突变群。直到最近，由于普遍关注物种的特异性生物化学、生理学或发育等方面的研究，大部分突变体被废除使用。然而最近几年，从单个基因研究向功能基因组研究的转变改变了这种观点，全基因组范围诱发突变项目现在正进行收集尽可能多的影响各种生物学过程的突变体的材料。最终的目的是建立一个广泛的突变体文库作为所有研究者的中心资源。例如，小鼠的第一个基因组范围 ENU 突变计划涉及惊人的 40 000 个小鼠品系医学相关表型的筛查，包括临床生物化学、突变反应、免疫学、对生理刺激的反应、发育缺陷和行为等方面（Hrabe de Angelis and Balling，1998，Hrabe de Angelis *et al*.，2000，Nolan *et al*.，2000）。这些项目正在进行并且其结果在表 20.3 列出的网址上不断更新。

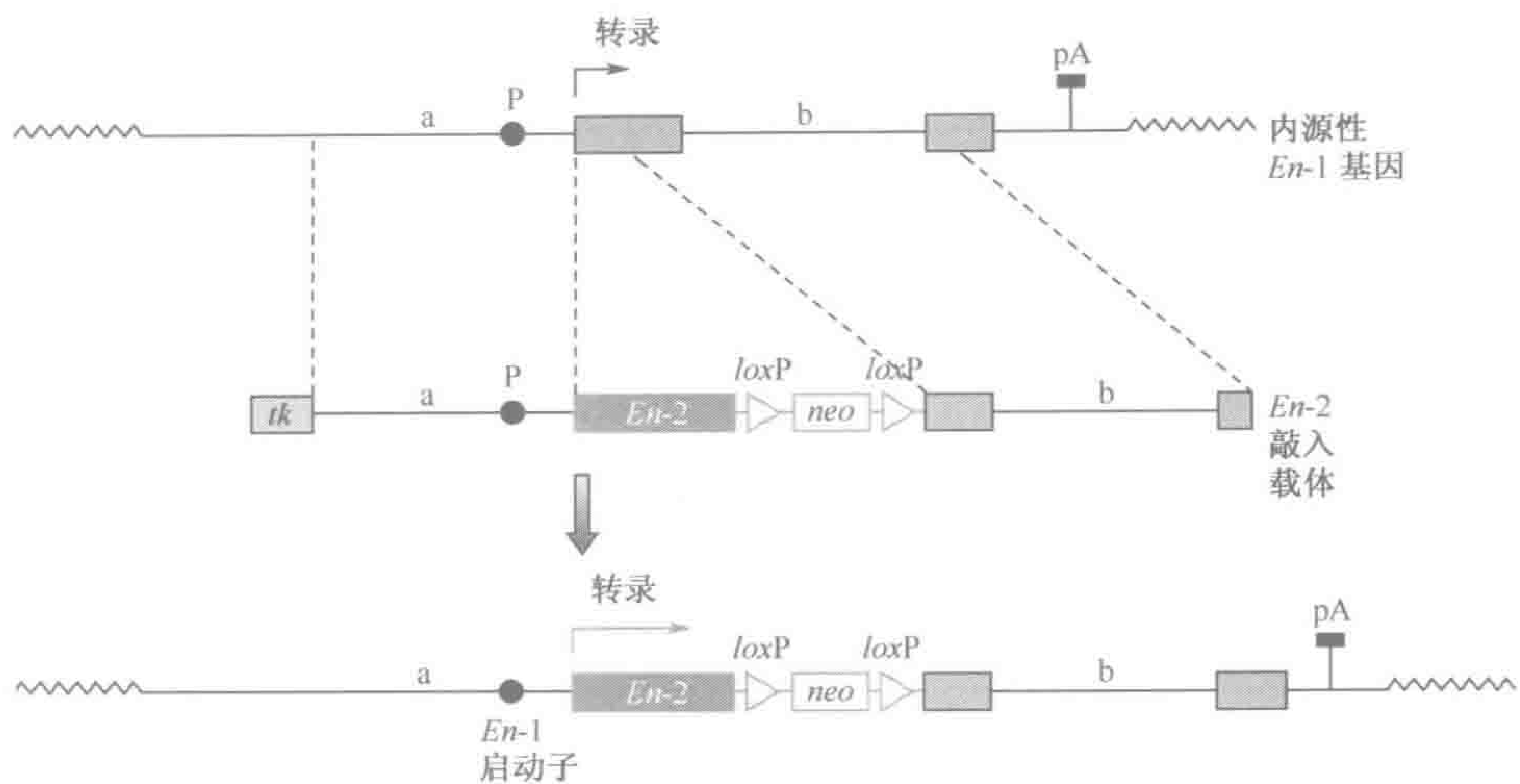


图 20.14 基因敲入方法通过导入一个基因来替代染色体基因的活性

最顶部显示 *En-1* 基因有两个外显子，编码序列用实充框表示。也标出启动子（P）和多聚腺苷酸位点（pA）。基因打靶载体（“knock-in 载体”）含有克隆的 *En-1* 基因序列，该序列由上游序列 a 和内部序列 b 组成。序列 a 包含 *En-1* 启动子，序列 b 包含外显子 1 的 3′端序列，单个内含子和外显子 2 的 5′端编码序列。将 a、b 这两个序列分开的是 *En-2* 基因的编码序列。两个标记匣包括一个胸苷激酶（*tk*）基因和一个新霉素耐药基因（*neo*），它们都受磷酸葡萄糖激酶启动子的控制（选择该启动子是因为磷酸葡萄糖激酶在 ES 细胞中表达）。*neo* 基因两侧有 *lox*P 序列，打靶方法导致内源序列 a 和 b 被替代而 *En-1* 基因的 5′端编码序列缺失。敲入的 *En-2* 基因受 *En-1* 启动子调控（Hanks *et al*. 1995）。注意：“敲入”一词已被应用在将新基因插入到内源基因并表达的任何方法，即使这后者只作为一个报道基因，诸如 *lacZ*。

化学诱变剂和放射线容易诱导点突变。尽管这些突变可能是“真实的”，它们代表引起人类疾病突变的类型，但主要的问题是对单个突变动物的 DNA 进行精细结构改变的确认，常常需要通过繁琐的位置克隆方法。可替代的是用基因转移作为诱发突变。在这种情况下突变剂是插入的 DNA 序列、以任何其他名称命名的一个转基因，偶尔会整合到已存在的基因中，并干扰它的表达。这种策略比放射线和化学诱变有一个主要优点，即在突变的基因座留有一序列标签。因此能快速鉴定突变基因座的分子特征（框 20.5）。在果蝇，P 元件用做插入突变剂并被直接导入生殖系（节 20.2.2）。ES 细胞提供了将这样突变导入小鼠的方法，因为 DNA 序列的随机整合在任何插入的一些基因都可被重新找到。然而，只有 3%的小鼠基因组是基因序列，大部分随机插入事件不会导致基因破坏。

为了提高重新找到插入基因的可能性，设计了**基因捕获**（gene trap）的方法（Evans *et al*.，1997）。其基本原理是转基因构件中含有一个有缺陷的报道基因或可筛选标记，只有当构件插入到一个基因内才可以表达，例如，转基因可以由一个上游剪切受体位点的报道基因组成。如果这个整合发生在基因内，它可以作为一个附加的外显子。当此被干扰的基因转录时，报道基因的转录物应被剪切到上游外显子上，当翻译时，保留它的报道子活性。这种方法的优点是，即使在杂合子状态，报道基因表达模式也与被破坏基因一样，那么，尽管被破坏基因在杂合子状态是致死的，但是仍然能够得到基因的

信息（它的表达谱）。缺点是报道基因的表达依赖于被扰乱基因的表达，因此如果被扰乱基因不表达，报道基因也不能表达。这个问题已用报道基因自身组成性的启动子的表达解决了，但其产生多聚腺苷酸作用仍然依赖于周围的基因（图 20.15）。以最近一篇描述小鼠胚胎干细胞的 2000 个基因的破坏和序列鉴定的报道说明这种方法已经被广泛应用。(Zambrowicz *et al.*，1998)。两个主要的创始小组正在进行小鼠的研究，一个是由德国 Gene Trap 协会组织的，目标是制作并研究 20000 个基因捕获品系（Wiles *et al.*，2000）；另一个是由美国 Lexicon Genetics 公司组织的。二者都将插入的侧翼序列保留在数据库中。这样研究人员可在实验中确定他们感兴趣的基因并对其进行检索，得到他们想要的基因失活的鼠系。果蝇的基因捕获计划也在进行（Berkeley Drosophila 基因组计划，见 Sparadling *et al.*，1995，1999），该计划不但包括 P 元件介导的基因干扰还包括 P 元件调节的基因激活，后一方法（激活示踪，框 20.5）用 P 元件整合一个强的面向外的启动子。一旦发生整合，这些构件便可激活相邻的内源基因。这样插入诱变可以产生功能获得突变以及功能丢失突变。所有上述计划的国际互联网资源都列入表 20.3。

框 20.5　用于插入诱变的复杂载体

任何 DNA 序列都可用作插入载体，并且只要是宿主基因组内部不存在的序列，就可以作为一个标签通过简单的杂交和 PCR 实验确认被破坏的基因。然而，包含特定的外部特点也可增加载体的功能，揭示被破坏基因及其产物的更多信息，甚至有助于插入位点周围的侧翼序列的克隆。

基因捕获（gene trap）　插入载体包括一个报道基因，诸如 *lacZ* 基因在剪切受体位点的下游，这样如果载体插入到一个基因内报道基因就表达。一个修改的基因捕获载体，有时称启动子捕获，只有当载体整合到基因中才表达，它含有一个无启动子的天然报道基因，只有当它插入到有活性启动子的下游才被激活（Evans *et al.*，1997）。这两种情况，报道基因表达都依赖于这个被破坏的基因。这种方法的好处是通过报道基因（提供功能信息）的表达反映了被破坏基因的表达模式，但缺点是插入到不表达的基因时报道基因就不能表达。用组成性启动子表达报道基因可克服这个缺点，但产生多聚腺苷酸作用仍依赖于被破坏的基因（Zambrowicz *et al.*，1998）。

增强子捕获（enhancer trap）　插入载体包括一个报道基因诸如 *lacZ*（通常是一个 TATA 盒），在一个小的启动子下游只支持自身的背景转录。当构件整合到内源性增强子所能影响的范围内时，报道基因模仿该增强子控制的基因表达情况（O'Kane and Gehring，1987）。由于增强子可以远距离发挥作用，这个策略很少用于基因的鉴定，但细胞特异性增强子可用于其他目的的研究，诸如控制 Cre 表达（节 20.5）或用于细胞消除的负性可筛选标记的激活（框 20.2）。

活化标签（activation tag）　插入载体含有一个强的向外面的启动子。如果此序列整合到某基因邻近的位置，该基因可被启动子激活，可以通过异位表达或过表达产生功能获得表型（Rorth *et al.*，1998）。

质粒拯救（plasmid rescue）　插入载体含有细菌的复制起点和抗生素耐药标记。这就意味着如果来自转基因插入的序列的基因组 DNA 用限制性内切酶消化、稀释、并用连接酶环化，此插入序列及其介导的旁侧序列将形成一个功能性质粒。如果用 *en masse* 转化细菌，收集整个基因组 DNA 环，则只含有这个质粒的细胞能够生存。这是一快速克隆和确认插入位点周围的基因序列的方法（Perucho *et al.*，1990）。

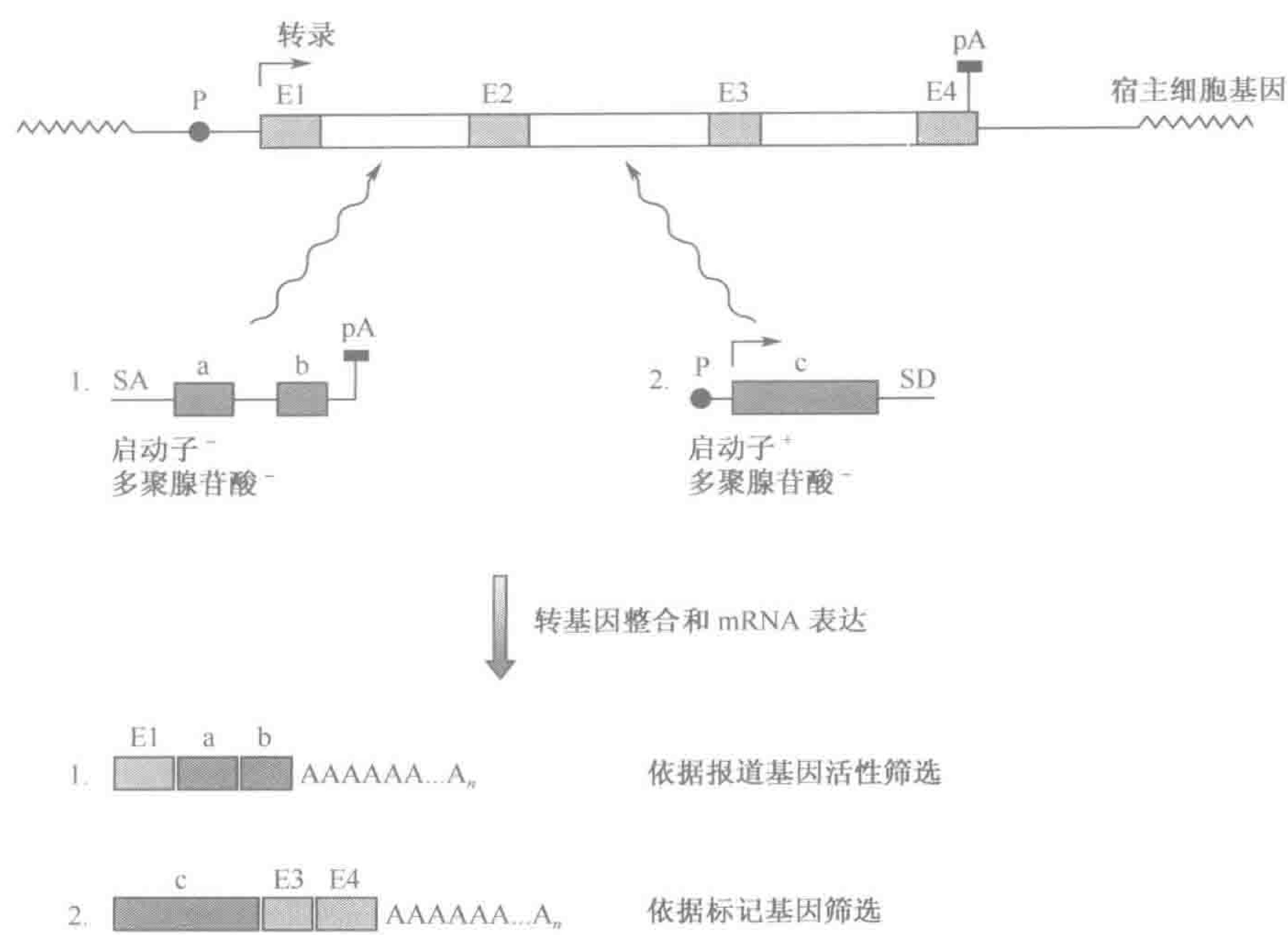


图 20.15　用表达缺陷的转基因选择发生在一个基因内或其附近的染色体整合事件的基因捕获
在顶部显示的是一个宿主细胞基因，它有四个外显子（E1～E4），一个启动子（P）和多聚腺苷酸化信号（pA），图中显示了两个可用于基因捕获的载体。转基因 1 包含一个缺乏启动子的报道基因。它有两个外显子序列 a 和 b，一个多聚腺苷酸化信号（pA）和一个含有**剪接受体序列**（splice acceptor sequence，SA）的上游序列。在这个例子中，转基因 1 整合到宿主基因的内含字子 1 上。在内源基因启动子转录过程中，剪接受体序列将有助于转基因外显子与内源基因的第一个外显子的拼接，产生一个融合的转录物。根据功能报道子的活性来筛选这个转录物（Evans *et al*.，1997）。转基因 2 有一个标记基因（耐药基因如嘌呤霉素 N-乙酰基转移酶），连接一个在 ES 细胞内工作的启动子（通常是磷酸葡萄糖激酶启动子）和一个下游**剪接供体序列**（splice donor sequence），但缺少多聚腺苷酸化信号。再者，整合的目的是允许表达，但最终还取决于转基因启动子和有助于 RNA 转录物剪接到宿主外显子下游的剪接供体序列。

RNA 干涉（节 20.2.6）是用于高通量基因功能注释的另外一种方法。迄今为止，**基因敲落**（gene knock-down）只在秀丽新小杆线虫中广泛应用，但它已显示出在哺乳动物细胞中应用的巨大潜力，或许只有此方法适合于快速而直接对人类基因的注释，至少它们的功能能在细胞水平确定。一些大规模的实验已经在线虫进行，涉及几千个 dsRNA 的合成并通过注射、浸泡或喂食的方法系统研究对线虫的影响（Gonczy *et al*.，2000。Maeda *et al*.，2001，Fraser *et al*.，2000）。最近，通过大量筛查得到将近17 000 个细菌菌株，每一个菌株表达一个不同的 dsRNA，它们是线虫基因组 86% 的基因（Kamath *et al*.，2003），这些线虫的 10% 以上显示了繁育致死、不育、生长或发育模拟表型。

20.4　利用基因转移和基因打靶技术建立疾病模型

人类疾病模型对医学研究非常重要。动物模型尤其可用于对疾病生理基础进行详细检测，它们为新的治疗方案在临床应用于人之前的疗效研究提供了新的前线测试系统。尽管细胞培养对于疾病模型的资源来说很有限，但它提供了一个替代方法，尤其是用来研究疾病的生化效应和药物反应。虽然许多个别的人类疾病没有一个好的动物模型，但一些主要人类疾病的代表性模型都已经建立：如遗传性疾病、感染性疾病、散发性癌和自身免疫性疾病等（leiter *et al*.，1997；Darling and Abbott，1992；Clarke，1994；Bedell *et al*.，1997）。一些人类疾病的动物模型是自发形成的；其他的是用不同的方法人工制作的（表 20.4）。迄今已经有很多自发性的或随机诱发的突变动物模型可供使用。最近基因打靶和转基因技术为动物或细胞疾病模型的制作提供了直接的方法，小鼠的靶向突变已经非常有用。有趣的是通过人和小鼠之间种间同源突变的比较越来越清楚地显示疾病的表型存在明显不同（节 20.4.6）。

表 20.4　动物模型分类

制备程序	举例	注释
(A)自发性的	生殖系突变→遗传疾病	
	体细胞突变→癌	
(B)人工干预或人工诱导产生的	通过筛选性繁育得到对疾病遗传性易感的品系	
	用相关的病原微生物感染的动物品系	
	通过操作环境诱导疾病而不产生突变	例如：将降植烷(pristane)，一种合成的辅助油，注射到大鼠皮下，制作关节炎模型
	用强突变剂在体内诱发突变，诸如 X 射线或有力的化学诱变剂如乙基亚硝苯脲(ENU)	已建立大规模化学诱发突变项目，制备突变小鼠和斑马鱼模型（节 20.4.2）
	遗传受精卵细胞或早期胚胎细胞的修饰及后期动物繁育（转基因和基因打靶技术）	（节 20.1～20.3）

20.4.1　在培养细胞中模拟疾病发生和药物治疗

疾病发生的许多方面都有可在体外进行有效研究的一个细胞成分，这样的例子包括 DNA 修复缺陷的着色性干皮病和范康尼贫血症，Werner 综合征的未成熟细胞衰老特征，见于神经纤维瘤的细胞骨架异常结构，当然还有失去增殖控制的癌症。只要这样的表型存在，细胞模型的开发就比动物模型更容易，因为将一个致病的显性基因转移到适合类型的细胞中很容易建立获得功能缺陷模型，而功能丢失模型目前可用 RNA 干涉方法有效地制备。细胞模型可基于人类细胞系，在许多情况下，不需要人工制备细胞模型，因为可以从相应疾病的个体分离到细胞。当然，动物更适合，因为它们提供了研究

无适当的细胞表型的疾病需要的一个整个生物体的背景（例如，自发性振颤）。然而很多疾病，细胞模型是首要的研究路线，它允许那些后期可用动物模型进行检测的有前景的候选药物的许多主要成分的筛查和鉴定。

20.4.2　自发性和随机诱变产生的动物疾病模型很难确认

自发性动物疾病模型

突变体人类表型，尤其是与明显疾病症状相关的表型正在进行认真地研究，多数罹患疾病的个体都去看医生，如果他们有新的早先没有报道过的表型，其病例可提供给专家，它们常在医学文献中记录此表型。依据受累的个体、他们的家庭、医生和感兴趣的医学研究人员以及做大规模群体筛查的工作人员等（全球群体超过 600 万个体）的不同目的，制定非常有效的筛查突变表型的方法。相比之下，许多动物疾病表型并未被记载。只有一小部分动物群是受控制的，且自发性突变表型的记录主要取决于为研究目的对繁育动物群落和对很少一部分的家畜及宠物的检测。只有明显外部异常的突变体才可能被注意。尽管鉴定自发性突变体很困难，但已经有很多类似于人类疾病的动物模型被描述（见表 20.5 的一些例子）。有的动物突变体表型和相应的临床表型非常相近，但由于在生化和发育通路上存在着差异，所以在其他方面还有相当的歧异（Erickson，1996；Wynshaw-Boris，1996）。另外，由于种间同源基因座的不同类突变会导致不同的表型差异。

表 20.5　自发性动物突变体的例子

动物突变体	表型特点和致病分子
NOD 小鼠	非肥胖型糖尿病，与人胰岛素依赖性糖尿病症相似
mdx 小鼠	由于小鼠肌萎缩蛋白基因突变导致的 X-连锁的肌营养不良，mdx 的原始突变是无义突变，但表型比在杜兴肌营养不良（DMD）的表型轻
血友病狗	凝血因子Ⅸ基因错义突变引起功能完全缺失，人的同源体是血友病 B
Watanabe 遗传性高血脂（WHHL）兔	低密度脂蛋白受体基因（LDLR）缺失四个密码子导致的高脂血症 与人的家族性高胆固醇血症相似
动脉粥样硬化猪	以高胆固醇为特点，LDL 受体活性正常，但有各种脂蛋白改变，包括脂蛋白 B
斑点小鼠	异常色素沉着，表型与 Waardenburg 综合征重叠，提示这种小鼠是 Waardenburg 综合征的动物模型，并通过人和小鼠同源 *PAX* 基因的鉴定而得到证实
NF-少女鱼	广泛的神经纤维瘤提示它可能是人纤维瘤Ⅰ型同源体

用化学诱变剂和发射线产生随机突变

产生动物突变体的传统方法涉及暴露于可控制的化学突变剂或高剂量 X-射线。通过这种方法获得了大量的果蝇和小鼠的突变体，并正如上述的，在小鼠和斑马鱼（具有动物模型的某些优点；框 20.6）已经进行了有效的大规模的突变筛查。然而，用化学诱变和放射线诱变问题的最主要是它们产生的突变基本上是随机的。为了确定感兴趣的

突变表型，在诱发突变后需要通过细致的表型检测对突变体进行辛苦的筛查。这些研究中发现的突变体表型，像自发性突变一样，与明显的外部异常表型有着偏差，只是因为容易简单地鉴定它们的表型。不过，人类疾病的若干重要模型已经用这种方法建立。

20.4.3 小鼠被广泛应用于人类疾病的动物模型主要是因为在预定的基因座建立特异的突变

在已经描述过具有不同潜力的许多动物物种中建立了自发性或人工诱导的人类疾病模型。无脊椎动物诸如果蝇和线虫，甚至酵母细胞都能够提供一些有用的动物模型，反映了特定蛋白质的存在和通路在进化过程中是非常保守的（节 3.9.2）。进化上距离远的脊椎动物如斑马鱼，也具有作为模式生物的一些优点，已用于某些人类疾病模型（框 20.6）。哺乳动物被认为应该能够提供更好的动物模型，但是由于各种原因，我们最近的亲属，类人猿，在提供疾病模型上没有什么用处。而其他的哺乳动物，尤其是小鼠，已经被广泛地用作人类疾病的模型（框 20.6）。在通过暴露于化学诱变剂和放射线人工诱导的动物模型或自发性产生的动物模型中，对产生的突变表型没有或很少有人为的对照，而对动物疾病模型的鉴定常常是偶然的。转基因/打靶小鼠疾病模型的巨大优势是按目的创建特定的疾病模型。如果相关的基因克隆是有用的，包括一些情况的突变基因，就可以制备想要选择的改变靶基因的小鼠。用这种方法可以建立所有主要类型疾病，遗传疾病、肿瘤、感染疾病和免疫缺陷疾病的模型（表 20.6；Smithies，1993，Clarke，1994；Bedell *et al.*，1997）。在大多数情况下，转基因/基因打靶方法用于单基因疾病的模型，但也正不断尝试用于制作复杂遗传病的小鼠模型，诸如阿尔海默氏病，动脉粥样硬化和原发性高血压（Petters and Sommer，2000）。

框 20.6 人类疾病模型动物的潜能

灵长类（primate） 因为它们与我们如此密切的关系，所以应该可提供最好的人类疾病动物模型（节 12.4.2 ）。人类和巨人猿在发育、解剖、生化和生理方面有很广泛的相似性。但灵长类的繁育昂贵，圈养的群体规模也很小，当然公众的敏感性也有一定的作用。尽管一些人原则上反对所有的动物实验，但从医学研究利益出发同意动物实验并认为用小型的实验动物诸如小鼠、大鼠更合适。更重要的是灵长类不适用于实验：它们的寿命比较长，并且比啮齿类繁育数量少，管理繁育实验很困难。尽管已有新的许多治疗人类疾病的方法正在快速开发起来（21 章），但由于灵长类检测实验需要太长时间所以促使了替代模型的研究。

小鼠 在人类疾病模型中应用最为广泛。它们体积小并且能相对经济地维持繁育。寿命短(2～3 年)，生殖周期短（～3 个月），繁殖量大（平均一个雌鼠能生 4～8 窝，每窝 6～8 只小鼠）。因为容易繁育，可选用复杂的繁育程序培育**重组杂交品系**（recombinant inbred strain）和**同类品系**（congenic strain）（Taylor，1989；框 14.2），因为繁育周期和寿命短，所以可以相对容易地监测到几代之间致病突变传递的效应。实验小鼠遗传学已被广泛地研究了几十年，许多突变体的表型都有记载（Lyon and Searle，1989）。大多数这样的突变体都是繁育群落中自发产生的。少数是人工诱变产生的，开始的时候使用 X-射线或化学突变剂制备突变，后来逐渐通过基因打靶和插入突变来构建。通过**种间回交定位**（interspecific back-cross mapping）和大量可使用的多态标记（～9000 二核苷酸重复标记已经被作图）小鼠的突变体作图是很方便的（Avner *et al.*，1988；框 14.2）。因

框 20.6　人类疾病模型动物的潜能（续）

为小鼠和人类基因组的同线性区域已经整理（图 12.11），这些信息在真正确定人和鼠同源性单基因病是很有用。

大鼠　相对比较大，在生理、药理和行为实验方面，尤其在心血管和神经精神的研究上更有意义。大鼠的生殖周期长（11 周），繁育较贵。一些人类疾病（例如高血压和行为疾病）还没有好的小鼠模型，正依赖于大鼠模型的建立。大鼠基因组遗传图和物理图已绘制完成，最近已经获得大鼠基因组序列（节 8.4.4）。

斑马鱼　是脊椎动物发育的理想模型，因为鱼在体外发育并且胚胎像青蛙的一样坚硬，生殖间隔和小鼠相似，但在相同时间可产生的后代比小鼠多 40 倍，并且斑马鱼的胚胎像果蝇和线虫一样是透明的。斑马鱼的遗传学研究是先进的：从 20 世纪 90 年代中期，对斑马鱼进行了大规模的突变筛查，基因转移技术是常规方法，密集的遗传图和物理图与 EST 资源都可供使用。斑马鱼和人类基因组有中等程度的同线性。在斑马鱼的遗传筛查中鉴定的许多突变表型与人的疾病状态相似，为人类疾病，尤其是血液疾病、心血管疾病和肾脏疾病提供了有用的疾病模型。例如，*sauternes* 基因是与人 *ALAS2* 同源的斑马鱼基因，编码 δ-氨基乙酰丙酸合成酶-亚铁血红素生物合成中的第一个酶。斑马鱼突变体提供了第一个遗传性铁粒幼红细胞贫血的动物模型，由 ALAS2 功能缺失引起（Brownline *et al.*，1998）。Dracula 基因是人 *FECH* 基因的斑马鱼种间同源体，编码亚铁整合酶-亚铁血红素合成过程中最后的一个酶。这个突变提供了一个红细胞生成原卟啉的模型（Childs *et al.*，2000）。最近已有关于其他模型的综述（Dooley and Zon，2000）。尽管斑马鱼和人类基因组表现了一定的相似性但由于他们有很大程度的进化上的不同，因此人类疾病和斑马鱼突变体的相关性被界定为影响高度保守通路的疾病。许多斑马鱼突变体已成为很好的人类疾病模型，并因此用于候选药物的检测。这些突变体包括阿尔茨海默氏症、亚铁血红素生物合成异常、先天性心脏病、多囊肾病和癌。

无脊椎动物和酵母是通过几百万年的进化从人类分离的，但一些核心基因和基本通路仍是保守的。所以，简单的有机体，诸如酵母 *saccahromyas cereviae* 和线虫 *caenorhabditis elegans* 提供某些疾病的有用模型，至少提供一药物实验系统，即使整体表型并不特别相应。例如，胰岛素通路在人类和线虫之间是完全保守的，所以线虫可用来作为糖尿病的模型。同样地，影响很基本的细胞功能（诸如在 Bloom 综合征中 DNA 螺旋酶的缺陷）的疾病能以酵母做模型。重要地是酵母和秀丽新杆小线虫能作为微生物处理，并因此能以引导化合物和候选药物用嵌板成批筛查。

表 20.6　人类疾病转基因或基因打靶小鼠模型举例

人类疾病或异常表型	基因	构建模型的方法
囊性纤维性变	*CFTR*	基因打靶插入失活
β-地中海贫血	*HBB*(β-血红蛋白)	基因打靶插入失活
高胆固醇血症和动脉粥样硬化	载脂蛋白基因，例如 *APOE*	基因打靶插入失活
Gaucher's 病	*GBA*	基因打靶插入失活
脆性一X 综合征	*FMR1*	基因打靶插入失活
GSS 综合征	朊病毒基因(*PRNP*)	突变的小鼠朊病毒基因整合
脊小脑共济失调 1 型(SCA1)	*SCA1*(共济失调)	突变的人共济失调基因和扩张性三核苷酸重复整合
阿尔茨海默氏症	*APP*(β 淀粉样前体蛋白)	突变的全长 *APP* cDNA 在血小板生长因子启动子调控下整合

20.4.4 基因打靶可用于制作功能丢失突变模型，而用显性突变基因表达制作获得功能突变的模型

通过基因打靶制备小鼠功能丢失模型

许多疾病表型，包括所有原发隐性遗传病和许多显性遗传病的表型，都认为是由于基因功能的丢失产生的。为这类单基因病建造疾病模型的最简单的方法是构建基因敲除小鼠。第一步是分离种间同源性小鼠的基因并用它的一个片段通过基因打靶技术敲除小鼠 ES 细胞的内源基因（节 20.2.4）。接着将遗传修饰的 ES 细胞注射到假孕小鼠的囊胚中继续发育，得到在其生殖细胞中带有相当大一部分靶向突变基因的**原代小鼠**（founder mice），这些小鼠可被杂交，其后代可以从尾部收集血液，用 PCR 方法筛查所期望的突变和野生型的等位基因。基因打靶事件是有意制备一个无效等位基因（那里基因表达完全缺如），但有的时候结果可能是**渗漏突变**（leaky mutation）和突变的等位基因仍然有一些表达。例如，由于异常剪切，跨越插入的 DNA 片段，可以得到一些正常基因的产物。这可以解释 Snouwaert 等（1992）和 Dorin 等（1992）制备的囊性纤维性变小鼠模型病情严重程度的不同。表型的不同也可能是由于使用不同品系的修饰基因造成的（节 20.4.6）。

利用表达显性突变基因制备功能获得突变模型

一般的实验设计经常结合基因转移的前核微注射技术。建造的疾病模型一定是导入的 DNA 就足以致病，包括遗传性获得功能突变和由原癌基因引起的散发性肿瘤。为了制备这些疾病模型，需要克隆一个突变基因，如果必要的话也可以通过体外诱发突变的方法设计一个突变。然后此突变基因作为一转基因简单地插入，例如通过微注射到受精卵的卵母细胞。因为不需在特定位点整合导入突变的基因，人类突变体的基因虽能够满足需要，但有时也用小鼠的突变体基因，下面的两个例子说明了这个方法。

- 一个早期的实验是想检测在 Germann-Sträussler-Scheinker（GSS）综合征患者中的朊病毒蛋白质基因第 102 个密码子被亮氨酸替代后是否致病。于是在克隆小鼠朊病毒中设计一个相同的突变，然后注射到受精的卵细胞中制备转基因小鼠，小鼠表现为自发性神经退行性变，重现人 GSS 综合征中的症状（Hsiao *et al.*，1990）。用朊病毒蛋白质转基因的各种实验对于朊病毒的认识都是很有帮助的（Gabizon and Taraboulos，1997）。
- 引起神经退行性病变的三核苷酸扩展性重复形成了另外一种类型的功能获得突变。脊髓小脑共济失调 I 型（SCA1）是共济失调基因三核苷酸 CAG 重复不稳定扩增的显性遗传疾病。以小脑浦肯野氏细胞、脊髓小脑束和一些脑干神经元退行性改变为特点。通过导入浦肯野氏细胞特异性启动子驱动两个转基因的一个：人的正常共济失调基因（SCA1）或含有扩展性 CAG 重复的突变共济失调基因制备转基因鼠，两种类型的转基因都表达，但只带有扩展性等位基因的小鼠发生共济失调、浦肯野氏细胞退行性改变，证实了功能获得的假设（Burright *et al.*，1995）。

用小鼠制备人类癌症模型

人们正致力于构建小鼠癌症模型（Ghebrainious and Donehower，1998；Macleod and Jacks，1999）

- 获得功能。在这种情况下，疾病是由于原癌基因不适当的激活所引起，可通过构建转基因小鼠来制备模型。通过简单的转基因整合将适当的原癌基因导入小鼠的基因组。
- 功能丢失。在这种情况下，疾病是由于肿瘤抑癌基因的失活所引起，通过打靶构建基因敲除小鼠模型。例如，有几个模型是使 *TP53* 和 *RB1* 的小鼠同源基因失活制备的，但此表型只显示分别与相应的人类 Li-Fraumeni 综合征和视网膜母细胞瘤表型大体相似（见 20.4.6 部分）。

20.4.5　用转基因动物制作复杂疾病的模型是日益关注的重点

制备染色体病模型

现存的人类染色体病的小鼠模型很少，有些情况是因为两个物种之间没有充分保守的同线性。以 Down 氏综合征（21 三体）为例，人类的 21 号染色体与小鼠的 16 号染色体有很大的同线性区域（图 12.11），但是 16 三体的小鼠不是很好的 Down 氏综合征的模型，因为它们在子宫内就死亡。的确，16 三体的小鼠也从来没被认为是人 21 三体的动物模型，因为这两个染色体不是完全相同的。人 21 号染色体远端 2～3Mb 区域的基因与小鼠的 17 号染色体和 10 号染色体具有种间同源性，小鼠 16 号染色体上的一些基因与人的其他染色体上的基因具有种间同源性，而非 21 号染色体。为了用小鼠制备更好的 Down 氏综合征模型，人们把注意力放在了 Down 氏综合征的关键区域，21q21.3-q22.2（通过观察少数含部分 21 三体的 Down 氏综合征患者表型推断而来）。在这个区域内，人的小脑基因（*minibrain*），位于 21q22.2，可能是与学习缺陷有关的重要基因座。含有 100kb 人小脑基因而显然不含其他基因的一个 180kb 的 YCA 转基因小鼠发育成学习缺陷小鼠（Smith *et al*.，1997）。为获得有学习和行为缺陷，由 Reeves 等（1995）用标准方法照射小鼠获得一部分 16 三体小鼠，Ts65Dn。Kola 和 Hertzog（1998）讨论了其他新近制备的一些模型。未来染色体疾病模型的制作设想利用 *Cre-lox*P 基因打靶的优点。正如在节 20.2.5 所描述的，这个系统为基因组工程提供了巨大的潜能，并可用于在预先选择的染色体上确定的位置进行染色体易位工程。YAC 转基因对基因过表达和由大范围基因剂量异常致病的其他染色体疾病的研究预计是很重要的，诸如 Charcot-Marie-Tooth 病 1A 型，就是由于 PMP22 基因区 1.5Mb 区域复制所致的过表达造成的（图 16.8）。

制备复杂疾病模型

人类遗传学的研究正转向对复杂疾病诸如动脉粥样硬化，原发性高血压和糖尿病等发病机制的探讨。这些疾病都有一个多遗传的和环境成分的复杂病因，虽然一些这样的疾病已经有了有价值的动物模型，但基因打靶技术有望在将来提供紧缺动物模型

(Smithies and Maeda, 1995)。只要确定了合适的致病相关基因，就可以通过繁育实验将致病基因一起进行不同组合，研究不同小鼠品系的不同遗传背景和不同环境因素的评估。这种方法并不是像听起来那么复杂，因为许多复杂疾病的表型逐渐被认为只是由于几个主要易感基因的组合。例如，隐性脊柱裂的双基因模型就是偶然的机会用杂合的 *Patch* 突变（*Pdgfrb*；血小板生长因子受体）小鼠和纯合波浪突变小鼠（*Pax-1*）交配繁育而得到的（Helwig *et al.*, 1995）。另一方面这种双基因模型提示可能的治疗策略。例如，*rds* 突变株小鼠（表现为视网膜退化）和表达 *Bcl-2*（抗细胞凋亡）转基因小鼠交配的导致后代视网膜退化减弱（Nir *et al.*, 2000）。

20.4.6 人和小鼠之间的各种差异使人类疾病的小鼠模型很难构建

自发性和人工诱发的动物模型的表型与相应的人类疾病有很多不同之处并非是不常见的。例如，利用基因打靶技术使小鼠的肿瘤抑制基因失活经常得到令人失望的动物模型，如像敲除 *TP53* 和 *RB1*（视网膜母细胞瘤）的情况。为想获得理想的突变可能存在一些问题。例如，上述讨论的敲除中的遗漏表达问题，或者在节 20.2.3 讨论的转基因的表达可能受各种因素的影响，诸如位置效应，导致意想不到的表型。除这些可能性外，小鼠与人在某些方面存在差异导致种间同源基因突变的分歧的疾病表型（Erickson, 1989, 1996; Wynshaw-Boris, 1996）。

- ▶ **生化通路的不同**　尽管哺乳动物的生化通路通常很保守，但小鼠和人的通路之间已知有所不同。人的视网膜极其依赖于 *RB1* 基因产物的精确功能，但其他脊椎动物的视网膜就不是这样，因此没有小鼠自发性视网膜母细胞瘤突变体，*Rb1* 基因敲除的小鼠也没有视网膜母细胞瘤的特征。但最近报道小鼠的正常胎盘发育需要 *Rb1*，*Rb1* 敲除小鼠表现为滋养层细胞过度增殖和严重的胎盘结构异常，导致血管减少，胎盘运输功能下降（Wu *et al.*, 2003）。另外一个例子是神经节苷脂退化通路。当人类 *HEXA* 基因突变（编码氨基己糖苷酶）时可导致严重的溶酶体储存异常的 Tay-Sachs 病，小鼠的种间同源 *HEXA* 基因失活导致神经细胞神经节苷脂异常积累，但没有人类的运动神经或学习缺陷（Wynshaw-Boris, 1996）。
- ▶ **发育通路的不同**　人和小鼠发育通路的差异还不十分清楚，但推测对一些重要的器官系统，诸如脑是非常重要的。
- ▶ **绝对时间**　由于小鼠和人的寿命极大的不同，在迟发的一些人类疾病可能很难制备小鼠模型。
- ▶ **遗传背景的不同**　这反映了修饰基因的重要性。大多数人群是远亲繁殖，然而实验用的鼠系是近亲繁殖。由于其他基因座［**修饰基因**（modifier gene）］的等位基因不同，它们可与感兴趣的基因座相互作用，因此在不同的鼠系表型可能非常不同。一个关于遗传背景重要性的有用实例是 *Min*（multiple intestinal neoplasia，**多发性肠息肉**）小鼠，它是由 ENU 诱发小鼠 Apc 基因突变产生的。人的种间同源 APC 基因突变引起腺瘤样息肉，与结肠癌有关，*Min* 小鼠被看作是该疾病很好的模型，然而 *Min* 小鼠表型受遗传背景修饰的影响，例如带有 APC_{Min} 小鼠结肠息肉的数量明显依赖于小鼠品系的不同。在人类同一家族的不同成员虽然都有相同的 *APC* 突变，但由于遗传背景的不同也具有不同的肿瘤表型。可能有时是由环境因素不同造成的，但

修饰基因的涉及一直很被怀疑。*Min* 小鼠提供了一个很好的用于定位和鉴定修饰基因的遗传系统（Dietrich *et al.*，1993；MacPhee *et al.*，1995）。

最近小鼠和大鼠基因组序列的完成已鉴定了许多在啮齿类没有其副本的人类基因，包括推测的常染色体 *SHOX* 基因（缺陷导致 Leri-Weill 综合征，也可能参与 Turner 综合征——见 Clement-Jones *et al.*，2000）和 Kalman 综合征基因，*KAL1*。

（郑志红　译）

进一步阅读

Kuhn R, Schwenk F (1997) Advances in gene targeting methods. *Curr. Opin. Immunol.* **9**, 183–188.

Popko B (ed.) (1998) *Mouse Models of Human Genetic Neurological Disease*. Plenum Press, New York.

Shastry BS (1998) Gene disruption in mice: models of development and disease. *Mol. Cell. Biochem.* **181**, 163–179.

Sikorski R, Peters R (1997) Transgenics on the internet. *Nature Biotechnol.* **15**, 289.

TBASE: a transgenic/targeted mutation database at http://www.gdb.org/Dan/tbase.Html.

Muller U (1999) Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis. *Mechanisms of Development* **82**, 3–21.

参考文献

Amaya E, Musci TJ, Kirschner MW (1991) Expression of a dominant negative mutant of the FGF receptor disrupts mesoderm formation in *Xenopus* embryos. *Cell* **66**, 257–270.

Avner P, Amar L, Dandolo L, Guenet JL (1988) Genetic-analysis of the mouse using interspecific crosses. *Trends Genet.* **4**, 18–23.

Bahramian MB, Zabl H (1999) Transcriptional and posttranscriptional silencing of rodent alpha1 (I) collagen by a homologous transcriptionally self-silenced transgene. *Mol. Cell. Biol.* **19**, 274–283.

Bedell MA, Jenkins NA, Copeland NG (1997) Mouse models of human disease. Part II. Recent progress and future directions. *Genes Dev.* **11**, 11–43.

Belshaw PJ, Ho SN, Crabtree GR, Schreiber SL (1996) Controlling protein association and subcellular localization with a synthetic ligand that induces heterodimerization of proteins. *Proc. Natl Acad. Sci. USA* **93**, 4604–4607.

Brownlie A, Donovan A, Pratt SJ *et al.* (1998) Positional cloning of the zebrafish sauternes gene: a model for congenital sideroblastic anaemia. *Nature Genet.* **20**, 244–250.

Burke AC (2000) *Hox* genes and the global patterning of the somite mesoderm. *Curr. Top. Dev. Biol.* **47**, 155–181.

Burright EN, Clark HB, Servadio A *et al.* (1995) SCA1 transgenic mice – a model for neurodegeneration caused by CAG trinucleotide expansion. *Cell* **82**, 937–948.

Campbell KHS, McWhir J, Richie WA, Wilmut I (1996) Sheep cloned by nuclear transfer from a cultured cell line. *Nature* **380**, 64–67.

Capecchi M (1989) The new mouse genetics: altering the genome by gene targeting. *Trends Genet.* **5**, 70–76.

Chan AWS, Homan EJ, Ballou LU, Bums JC, Brennel RD (1998) Transgenic cattle produced by reverse transcribed gene transfer in oocytes. *Proc. Natl Acad. Sci. USA* **95**, 14028–14033.

Chan AWS, Chong KY, Martinovich C, Simerly C, Shatten G (2001) Transgenic monkeys produced by retroviral gene transfer into mature oocytes. *Science* **291**, 309–312.

Childs S, Weinstein BM. Mohideen M-A PK *et al.* (2000) Zebrafish *dracula* encodes ferrochelatase and its mutation provides a model for erythropoietic protoporphyria. *Current Biol.* **10**, 1001–1004.

Clarke AR (1994) Murine genetic models of human disease. *Curr. Opin. Genet. Dev.* **4**, 453–460.

Clement-Jones M, Schiller S, Rao E, Blaschke RJ, Zuniga A, Zeller R *et al.* (2000) The short stature homeobox gene *SHOX* is involved in skeletal abnormalities in Turner syndrome. *Human Molecular Genetics* **9**, 695–702.

Dai Y, Vaught TD, Boone J, Chen S-H *et al.* (2002) Targeted disruption of the α-1,3-galactosyltransferase gene in cloned pigs. *Nature Biotechnol.* **20**, 251–255.

Darling SM, Abbott CM (1992) Mouse models of human single gene disorders. 1. Nontransgenic mice. *BioEssays*, **14**, 359–366.

De Wet JR, Wood KV, De Luca M, Helsinki DR, Subramani S (1987) Firefly luciferase gene: structure and expression in mammalian cells. *Mol. Cell. Biol.* **7**, 725–737.

Dietrich WF, Lander ES, Smith JS *et al.* (1993) Genetic identification of mom-1, a major modifier locus affecting min-induced intestinal neoplasia in the mouse. *Cell* **75**, 631–639.

Dooley K, Zon LI (2000) Zebrafish: a model system for the study of human disease. *Current Opin. Genet. Dev.* **10**, 252–256.

Dorin JR, Dickinson P, Alton EW *et al.* (1992) Cystic-fibrosis in the mouse by targeted insertional mutagenesis. *Nature* **359**, 211–215.

Efrat S, Lieser M, Wu Y *et al.* (1994) Ribozyme-mediated attenuation of pancreatic β-cell glucokinase expression in transgenic mice results in impaired glucose-induced insulin secretion. *Proc. Natl Acad. Sci. USA* **91**, 2051–2055.

Erickson RP (1989) Why isn't a mouse more like a man? *Trends*

Genet. **5**, 1–3.

Erickson RP (1996) Mouse models of human genetic disease: which mouse is more like a man? *Bioessays* **18**, 993–998.

Evans MJ, Kaufman MH (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156.

Evans MJ, Carlton MBL, Russ AP (1997) Gene trapping and functional genomics. *Trends Genet.* **13**, 370–374.

Famulok M, Verma S (2002) In vivo-applied functional RNAs as tools in proteomics and genomics research. *Trends Biotechnol.* **20**, 462–466.

Famulok M *et al.* (2001) Intramers as promising new tools in functional proteomics. *Chem. Biol.* **8**, 931–939.

Fiel R, Brocard J, Mascrez B, LeMur M, Metzger D, Chambon P (1996) Ligand-activated site-specific recombination in mice. *Proc. Natl Acad. Sci. USA* **93**, 10887–10890.

Figueiredo MS, Brownlee GG (1995) Cis-acting elements and transcription factors involved in the promoter activity of the human factor VIII gene. *J. Biol. Chem.* **270**, 11828–11838.

Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC (1998) Potent and specific genetic interference by double stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811.

Forss-Petter S, Danielsen PE, Catsicas S, Battenberg E, Price J, Nerenberg M, Sutcliffe JG (1990) Transgenic mice expressing β-galactoisidase in mature neurons under neuron-specific enolase promoter control. *Neuron* **5** (2), 187–197.

Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408**, 325–330.

Gabizon R, Taraboulos A (1997) Of mice and (mad) cows – transgenic mice help to understand prions. *Trends Genet.* **13**, 264–269.

Garrus JE, von Schwedler UK, Pornillos OW *et al.* (2001) Tsg101 and the vacuolar protein sorting pathway are essential for HIV-1 budding. *Cell* **107**, 55–65.

Ghebranious N, Donehower LA (1998) Mouse models in tumor suppression. *Oncogene* **17**, 3385–3400.

Gonczy P, Echerverri G, Oegema K, Coulson A *et al.* (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408**, 331–336.

Gordon JW (1992) Production of transgenic mice. *Methods. Enzymol.* **225**, 747–771.

Gorman CM, Moffat LF, Howard BH (1982) Recombinant genome which expresses chloramphenicol acetyltransferase in mammalian cells. *Mol. Cell. Biol.* **2**, 1044–1051.

Gu H, Marth JD, Orban PC, Mossmann H, Rajewsky K (1994) Deletion of a DNA-polymerase-beta gene segment in T-cells using cell-type-specific gene targeting. *Science* **265**, 103–107.

Hall CV, Jacob PE, Ringold GM, Lee F (1983) Expression and regulation of *Escherichia coli lacZ* gene fusions in mammalian cells. *J. Mol. Appl. Genet.* **2**, 101–109.

Hanks M, Wurst W, Anson-Cartwright L, Auerbach AB, Joyner AL (1995) Rescue of the *En-1* mutant phenotype by replacement of *En-1* with *En-2*. *Science* **269**, 679–682.

Helwig U, Imai K, Schmahl W, Thomas BE, Varnum DS, Nadeau JH, Balling R (1995) Interaction between *undulated* and *patch* leads to an extreme form of spina-bifida in double-mutant mice. *Nature Genet.* **11**, 60–63.

Herault Y, Rassoulzadegan M, Cuzin F, Duboule D (1998) Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nature Genet.* **20**, 381–384.

Higgins DE, Portnoy DA (1998) Bacterial delivery of DNA evolves. *Nature Biotechnol.* **16**, 138–139.

Hrabe de Angelis M, Balling R (1998) Large scale ENU screens in the mouse. Genetics meets genomics. *Mutations Res.* **400**, 25–32.

Hrabe de Angelis M, Flaswinkel H, Fuchs H, Rathkolb B *et al.* (2000) Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genet.* **25**, 444–447.

Hsiao KK, Scott M, Foster D, Groth DF, Dearmond SJ, Prusiner SB (1990) Spontaneous neurodegeneration in transgenic mice with mutant prion protein. *Science* **250**, 1587–1590.

Ikawa M, Yamada S, Nakanishi T, Okabe M (1999) Green fluorescent protein as a vital marker in mammals. *Curr. Top. Dev. Biol.* **44**, 1–20.

Jaenisch R, Mintz B (1974) Simian virus 40 DNA in DNA of healthy adult mice derived from preimplantation blastocysts injected with viral DNA. *Proc. Natl Acad. Sci. USA* **71**, 1250–1254.

Jakobovits A, Moore AL, Green LL *et al.* (1993) Germ-line transmission and expression of a human-derived yeast artificial chromosome. *Nature* **362**, 255–258.

Kamath RS, Fraser AG, Dong Y *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237.

Katsuki M, Sato M, Kimura M, Yokoyama M, Kobayashi K, Nomura T (1988) Conversion of normal behaviour to *shiverer* by myelin basic protein antisense cDNA in transgenic mice. *Science* **241**, 593–595.

Kola I, Hertzog PJ (1998) Down syndrome and mouse models. *Curr. Opin. Genet. Dev.* **8**, 316–321.

Koopman P, Gubbay J, Vivian N, Goodfellow P, Lovell-Badge R (1991) Male development of chromosomally female mice transgenic for *Sry*. *Nature* **351**, 117–121.

Kroll KL, Amaya E (1996) Transgenic *Xenopus* embryos from sperm nuclear transplantations reveal FGF signaling requirements during gastrulation. *Development* **122**, 3173–3183.

Kunik T, Tzfira T, Kapulnik Y, Gafni Y *et al.* (2001) Genetic transformation of HeLa cells by *Agrobacterium*. *Proc. Natl Acad. Sci. USA* **98**, 1871–1876.

Kuwabara T, Warashina M, Taira K (2000) Allosterically controllable maxizymes cleave mRNA with high efficiency and specificity. *Trends Biotechnol.* **18**, 462–468.

Lai L, Kolber-Simonds D, Park K-W, Cheong H-T *et al.* (2002) Production of α-1,3-galactosyltransferase knockout pigs by nuclear transfer cloning. *Science* **295**, 1089–1092.

Lamb BT, Gearhart JD (1995) YAC transgenics and the study of genetics and human disease. *Curr. Opin. Genet. Dev.* **5**, 342–348.

Leiter EH, Beamer WG, Shultz LD, Barker JE, Lane PW (1987) Mouse models of genetic diseases. *Birth Defects* **23**, 221–257.

Littlewood TD, Hancock DC, Danielian PS, Parker MG, Evan GI (1995) A modified oestrogen receptor ligand-binding domain as an improved switch for the regulation of heterologous proteins. *Nucleic Acids Res.* **23**, 1686–1690.

Lobe CG, Nagy A (1998) Conditional genome alteration in mice. *Bioessays* **20**, 200–208.

Lyon MF, Searle AG (1989) Genetic Variants and Strains of the Laboratory Mouse, 2nd edn. Oxford University Press, Oxford.

Macleod KF, Jacks T (1999) Insights into cancer from transgenic mouse models. *J. Pathol.* **187**, 43–60.

MacPhee M, Chepenik KP, Liddell RA, Nelson KK, Siracusa LD, Buchberg AM (1995) The secretory phospholipase-a2 gene is a candidate for the mom1 locus, a major modifier of apc(min)-induced intestinal neoplasia. *Cell* **81**, 957–966.

Maeda I, Kohara Y, Yamamoto M, Sugimoto A (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* **11**, 171–176.

Martin GR (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl Acad. Sci. USA* **78**, 7634–7638.

McCreath KJ, Howcroft J, Campbell KHS, Colman A, Schnieke AE, Kind AJ (2000) Production of gene targeted sheep by nuclear transfer from cultured somatic cells. *Nature* **405**, 1066–1069.

Melton DW (1994) Gene targeting in the mouse. *BioEssays* **16**, 633–638.

Mendez MJ, Green LL, Corvalan JRF *et al.* (1997) Functional transplant of megabase human immunoglobulin loci recapitulates human antibody response in mice. *Nature Genet.* **15**, 146–156.

Nir I, Kedzierski W, Chen J, Travis GH (2000) Expression of Bcl protects against photoreceptor degeneration in retinal degeneration slow (rds) mice. *J Neurosci.* **20**, 2150–2154.

Nolan PM, Peters J, Strivens M, Rogers D *et al.* (2000) A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet.* **25**, 440–443.

O'Kane CJ, Gehring WJ (1987) Detection in situ of genetic regulatory elements in *Drosophila*. *Proc. Natl Acad. Sci. USA* **84**, 9123–9127.

Pasquinelli AE (2002) MicroRNAs: deviants no longer. *Trends Genet.* **18**, 171–173.

Perry ACF, Wakayama T, Kishikawa H, Kasai T, Okabe M, Toyoda Y, Yanagimachi R (1999) Mammalian transgenesis by intracytoplasmic sperm injection. *Science* **284**, 1180–1183.

Perucho M, Hanahan D, Lipsich L, Wigler M (1980) Isolation of the chicken thymidine kinase gene by plasmid rescue. *Nature* **285**, 201–210.

Petters RM, Sommer JR (2000) Transgenic animals as models for human disease. *Transgenic Res.* **9**, 347–351.

Ramirez-Solis R, Liu P, Bradley A (1995) Chromosome engineering in mice. *Nature* **378**, 720–724.

Reeves RH, Irving NG, Moran TH *et al.* (1995) A mouse model for Down syndrome exhibits learning and behaviour deficits. *Nature Genet.* **11**, 177–183.

Richardson JH, Marasco WA (1995) Intracellular antibodies: development and therapeutic potential. *Trends Biotechnol.* **13**, 306–310.

Rorth P, Szabo K, Bailey A *et al.* (1998) Systematic gain-of-function genetics in *Drosophila*. *Development* **125**, 1049–1057.

Rudnicki MA, Braun B, Hinuma S, Jaenisch R (1992) Inactivation of *myoD* in mice leads to upregulation of the myogenic HLH gene *myf-5* and results in apparently normal muscle development. *Nucleic Acids Res.* **18**, 4833–4842.

Saez E, No D, West A, Evans RM (1997) Inducible gene expression in mammalian cells and transgenic mice. *Curr. Opin. Biotechnol.* **8**, 608–616.

Sakimura K, Kushiya E, Ogura A, Kudo Y, Katagiri T, Takahashi Y (1995) Upstream and intron regulatory regions for expression of the rat neuron-specific enolase gene. *Mol. Brain Res.* **28**, 19–28.

Schnieke AE, Kind AJ, Ritchie WA, Mycock K, Scott AR, Ritchie M, Wilmut I, Colman A, Campbell KH (1997) Human factor IX transgenic sheep produced by transfer of nuclei from transfected fetal fibroblasts. *Science* **278**, 2130–2133.

Shamblott MJ, Axelman J, Wang S *et al.* (1998) Derivation of pluripotent stem cells from cultured human primordial germ cells. *Proc. Natl Acad. Sci. USA*, **95**, 13726–13731.

Sklar MD, Thompson E, Welsh MJ *et al.* (1991) Depletion of c-*myc* with specific antisense sequences reverses the transformed phenotype in *ras* oncogene-transformed NIH 3T3 cells. *Mol. Cell. Biol.* **11**, 3699–3710.

Smith AJH, De Sousa MA, Kwabi-Addo B, Heppell-Parton A, Impey H, Rabbitts P (1995) A site-directed chromosomal translocation induced in embryonic stem-cells by Cre-*loxP* recombination. *Nature Genet.* **9**, 376–385.

Smith DJ, Stevens ME, Suclanagunta SP *et al.* (1997) Functional screening of 2 Mb of human chromosome 21q22.2 in transgenic mice implicates *minibrain* in learning defects associated with Down syndrome. *Nature Genet.* **16**, 28–36.

Smithies O, Gregg RG, Boggs SS *et al.* (1985) Insertion of DNA sequences into the human β-globin locus by homologous recombination. *Nature* **317**, 230–234.

Smithies O (1993) Animal models of human genetic diseases. *Trends Genet.* **9**, 112–116.

Smithies O, Maeda N (1995) Gene targeting approaches to complex genetic diseases: Atherosclerosis and essential hypertension. *Proc. Natl Acad. Sci. USA*, **92**, 5266–5272.

Snouwaert JN, Brigman KK, Latour AM *et al.* (1992) An animal-model for cystic-fibrosis made by gene targeting. *Science* **257**, 1083–1088.

Solter D, Gearhart J (1999) Putting stem cells to work. *Science* **283**, 1468–1470.

Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM (1995) Gene disruptions using P transposable elements: an integral component of the *Drosophila* genome project. *Proc. Natl Acad. Sci. USA* **92**, 10824–10830.

Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, Mozden N, Misra S, Rubin GM (1999) The BDGP Gene Disruption Project: single P element insertions mutating 25% of vital *Drosophila* genes. *Genetics* **153**, 135–177.

Szybalska EH, Szybalski E (1962) Genetics of human cell lines IV. DNA-mediated heritable transformation of a biochemical trait. *Proc. Natl Acad. Sci. USA* **48**, 2026–2031.

Taylor BA (1989) In: *Genetic Variants and Strains of the Laboratory Mouse*, 2nd edn (eds MF Lyon, AG Searle). Oxford University Press, Oxford, pp. 773–796.

Terskikh A, Fradkov A, Ermakova G *et al.* (2000) 'Fluorescent timer': protein that changes colour with time. *Science* **290**, 1585–1588.

Thomson JA, Itskovitz-Elder J, Shapiro SS *et al.* (1998) Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147.

Tomizuka K, Yoshida H, Uejima H *et al.* (1997) Functional expression and germline transmission of a human chromosome fragment in chimaeric mice. *Nature Genet.* **16**, 133–143.

Tuschl T, Borkhardt A (2002) Small interfering RNAs: A revolutionary tool for the analysis of gene function and gene therapy. *Molecular Interventions* **2**, 158–167.

Voinnet O (2002) RNA silencing: small RNAs as ubiquitous regulators of gene expression. *Current Opin. Plant Biol.* **5**, 444–451.

Wakayama T, Perry AC, Zuccotti M, Johnson KR, Yanagimachi R (1998) Full-term development of mice from enucleated oocytes injected with cumulus cell nuclei. *Nature* **394**, 369–374.

Welch PJ, Barber JR, Wong-Staal F (1998) Expression of ribozymes in gene transfer systems to modulate target RNA levels. *Curr. Opin. Biotechnol.* **9**, 486–496.

Wiles MV, Vauti F, Otte J *et al.* (2000) Establishment of a gene-trap sequence tag library to generate mutant mice from embryonic stem cells. *Nature Genet.* **24**, 13–14.

Wilmut I, Schnieke AE, McWhir J, Kind AJ, Campbell KHS (1997) Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813.

Wu LZ, de Bruin A, Saavedra HI, Starovic M, Trimboli A, Yang Y *et al.* (2003) Extra-embryonic function of Rb is essential for embryonic development and viability. *Nature* **421**, 942–947.

Wynshaw-Boris A (1996) Model mice and human disease. *Nature Genet.* **13**, 259–260.

Yokoyama M (2002) Gene delivery using temperature-responsive polymeric carriers. *Drug Discovery Today* **7**, 426–432.

Zambrowicz BP, Friedrich GA, Buxton EC, Lilleberg SL, Person C, Sands AT (1998) Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392**, 608–611.

第 21 章　疾病治疗的新方法

本 章 内 容

21.1　遗传病的治疗不同于疾病的遗传治疗

到了本书的结尾，关于治疗的章节理应考虑两个完全独立的事件：

- 遗传病的治疗；
- 疾病的遗传治疗。

后面的题目依次包括几个主要方面：

- 个体基因分型预测他们对药物治疗的良性反应及副作用；
- 利用遗传学和细胞生物学的知识识别药物发展的新靶标；
- 利用遗传学和细胞生物学的知识开发以细胞为基础的治疗方法；
- 利用遗传学技术生产药物、疫苗等用于疾病的治疗；
- 利用遗传学技术直接治疗疾病。

这些题目覆盖了 21 世纪大多数的医学研究。为了公正的评判如此大的领域需要由许多专家写出几本巨著，我们将在集中于利用遗传学技术治疗疾病的最后一个题目之前，对列表中前几个题目仅予以相当粗略的考查。这些发展带来的一些重要的伦理问题将在三个伦理学框中予以简要介绍。

21.2　遗传病的治疗

一般人的误解是若一种疾病是遗传的，它必定是不可治疗的。实际上疾病的病因和可治性间根本没有联系。一个深度耳聋的儿童，只有依据其症状及家庭条件应该给予听力帮助或者耳蜗移植，完全不考虑听力丧失是否是遗传的。传统的医疗以减轻疾病症状为目标，正像适合于任何其他疾病一样同样适用于遗传病。

然而，对许多遗传病来说现有的治疗确实不令人满意。20 年前的一份调查（Costa *et al.*，1983）估计，通过治疗，仅有 11%的孟德尔遗传病提高了生育能力，仅有 6%提高了社会适应能力，仅 15%（在那些寿命减少的患者中）的生命延长到正常水平。毫无疑问这些数据目前会有所提高，但并非显著改变。

先天性代谢缺陷是孟德尔疾病中适于传统治疗的最好的候选类型。我们详尽的生物化学知识为进行治疗提供了可能的切入点，包括：

▶ **底物限制**。众所周知苯丙酮尿症饮食治疗的成功（尽管在成功治疗的患者中认知的发育平均低于正常半个标准差）。同样其他几种出生缺陷对饮食治疗也反应良好；

▶ **缺陷产物的替代**。诸如利用甲状腺激素治疗先天性甲状腺机能低下患儿；

▶ **利用旁路途径去除有毒代谢物**。治疗措施包括从利用简单的放血疗法有效治疗血色素沉着症到利用苯甲酸盐提高尿素循环障碍患者氮的排泄；

▶ **使用代谢性抑制剂**。例如药物 NTBC 可以阻断酪氨酸代谢途径，并显著提高 1 型酪氨酸血症的预后（Lindstedt *et al.*，1992）。

Treacy，Valle 和 Scriver 详尽地阐述了这些和许多其他的例子，他们的章节会推荐给对此领域感兴趣的读者（进一步阅读）。事实上尽管做了许多工作，但几乎没有遗传病具有非常满意的治疗措施，因此下面着重阐述了同样适用于遗传病以及非遗传病的新方法。

21.3　利用遗传学知识改善现有治疗和发展传统治疗的新形式

21.3.1　药物遗传学有望提高药物的疗效并减少危险的副作用

罕见有药物对 100%的给予处方的患者有效。例如，在最常见的处方类药物中：

▶ 15%～35%的患者对 β 阻滞剂反应不充分或无反应；

▶ 7%～28%的患者对血管紧张素转换酶抑制剂反应不充分或无反应；

▶ 9%～23%的患者对选择性五羟色胺重摄取抑制剂反应不充分；

▶ 20%～50%的患者对三环类抗抑郁药反应不充分。

在某些病例（特别精神疾病），临床名称相同的患者实际上可能有不同的疾病，其中仅一种疾病对给予的药物有反应。但对于大部分，这些个体反应的差异主要依赖于药物吸收、分布、代谢和清除的差异以及靶受体的可变性。

这些个体差异主要依赖于有限数目基因上常见多态性的组合。与寻找常见疾病遗传易感因素的问题相比（15 章），鉴定构成个体对药物反应基础的变异体应该是一个更容易处理的问题。研究可集中于一个可确定的生化领域，一些假说经得起体外实验的检

验。因此，看起来能够实现个体特定处方的理想。某类以芯片为基础的试剂盒将用于临床对患者的几百个常见多态进行基因型分析，结果将决定在某病例中哪一种药物是安全有效的。愤世嫉俗者可能争辩说药品公司对确保他们的药物仅在哪些药物会发挥作用的病例中使用并不感兴趣，但对当药物可能引起危险的副作用时避免使用此药物很感兴趣。发展一种化合物从开始到市场的药，平均要花费 80 亿美元，历时 10～15 年，而且一些有前途的化合物因在少部分患者中的副作用而在后期被淘汰。已有几个例子可以说明影响药物反应的遗传变异体（表 21.1)。这些影响有实际的临床重要性。适于快乙酰化的异烟肼剂量在慢乙酰化中具有引起周围神经病变的风险。*CYP2C9* 基因有 R144C 或 I359L 变异的患者若给予标准维持剂量的华法林就会遭受极度抗凝及出血。*CYP2D6*“不良代谢变化者”可能对含有可待因的止疼剂无效，但如果给予 nortryptilene 可能每日只需 10～20mg，而超快代谢变化者每日可能需要 500 mg。Wolf 等（2000）引证了一个报道，在给予 CYP2D6 底物类精神药物的患者中，每一个具有使 *CYP2D6* 基因失活突变的患者都有药物的副反应。

表 21.1　遗传变异影响药物反应的例子

详见 Wolf 等(2000)

酶/蛋白	变异体	群体频率	受影响药物的例子
氮乙酰转移酶	慢乙酰化	60%(欧洲白种人) 20%(东方人)	异烟肼,普鲁卡因胺,磺胺药物
硫代嘌呤甲基转移酶	低活性	(低)	6-巯基嘌呤,咪唑硫嘌呤
硫代嘌呤甲基转移酶细胞色素 CYP2D6 P450	低活性	6%(欧洲白种人)	不能活化
	(失活突变)	1%(东方人)	可待因
	超高活性(串联基因扩增)	2%～7%(欧洲白种人)	慢失活:精神疾病类药物,例如 nortryptilene、氯氮平、氟哌啶醇、丙咪嗪、米安色林;心血管药物,如普萘洛尔
细胞色素 CYP2C9 P450	低活性	0.2%?	布洛芬、华法林、甲苯磺丁脲
细胞色素 CYP2C19 P450	低活性	4%(欧洲白种人) 23%(东方人)	3-甲基苯乙妥因、白乐君
β-肾上腺素能受体 ADRB2	几个 SNP—还不清楚哪些重要	NA	在哮喘中对沙丁胺醇的不同反应
ERBB2 受体	在某些乳腺癌中扩增	—	Herceptin(仅当 ERBB2 扩增时有效)

NA,无资料

21.3.2　药品公司投重资于基因组学以试图鉴定新的药物靶标

据说目前市场上药物全部的多样性是通过仅约 400 个靶标起作用。基因组学研究使药品公司可用于研究的潜在的靶标的数目已经从 10 年前的几百个大量的扩充至现在约 50 000 个。实际上，潜在靶标的过度富裕对于生物技术公司几乎是一个困惑。在投资鉴定可能抑制子所必需的大规模的筛查前，RNA 干扰（节 20.2.6）可用于鉴定抑制性药物可能产生有益作用的基因。从长远看来，通过这一努力有希望出现全新的药物类型。

作为新疫苗或治疗方法开发的指南对病原微生物的基因组学和蛋白质组学研究也很重要。微生物病原体被优先列入基因组测序（框 8.8），最近一个备受关注的成功例子为疟原虫的基因组测序（Gardner *et al.*，2002）。序列分析用来鉴定病原体而非宿主的特异的酶。这些酶可能是抑制子的靶标，或者是供应必须营养时使寄生菌易受干扰丢失的酶。感染早期表达的有毒蛋白质或基因产物是有前景的药物靶标。可以通过表达阵列研究以及比较有毒菌株和无毒菌株的基因含量和基因表达来鉴别。蛋白质组学策略（诸如 MALDI-TOF MS；框 19.4）用于鉴定病原体细胞表面可能是疫苗靶标的蛋白质。

21.3.3　基于细胞的治疗有望改变移植潜力

基于细胞的治疗可看作是目前移植方法的自然延伸，但是随着我们对干细胞了解的深入有望迅速扩大目前选择的范围。理论可能性是无止境的，使用适宜干细胞，任何受损的和衰竭的组织和器官都可以修复或重造，而这些干细胞可能预先经受任何类型的基因操作。Daley（2002）对这些设想多久能在实验室中实现提出了审慎的估计。

干细胞可定义为能够自我更新并可产生分化后代的细胞（节 3.3），它们可能存在于所有发育阶段及所有组织。从能够潜在地分化为生物体全部生殖系和体细胞的胚胎干细胞（ES 细胞）到多潜能的但有组织限制性的干细胞是有梯度的（图 21.1）。人们研究小鼠 ES 细胞已有多年，而 Thomson 等于 1998 年完成人 ES 细胞的分离。将干细胞治疗推到了医学研究和伦理争论的前沿。伦理争论源于产生干细胞的方法，该方法不可避免地涉及破坏早期人胚胎。如对这些讨论有兴趣，可参阅 Weissman 的综述［*New England Journal of Medicine* 346，1619～1622（2002）］。干细胞应用的反对者通常认为组织限制性的成体干细胞可以同样很好地发挥作用，因此应用干细胞是不必要的。另一些人则主张可获得的成体干细胞中几乎没有明确的特点，而且其生长和分化的潜能太有限，以至于不能适当的替代 ES 细胞。

目前的细胞系在培养基中的生长状况和产生的分化细胞类别上有很大不同。一个未解决的问题是组织限制性的干细胞通过迁移到不同组织并呈现该组织干细胞特性而转分化达到何种程度。现已有许多主张，但几乎没有被严格确定的。作为研究工具所有类型的干细胞都是非常重要的，它们可用来识别控制细胞是否分化和如何分化的因子。如果有能力控制和引导分化，干细胞则可能用于组织工程和组织修复的无限领域。

利用取自供体的干细胞仍留有移植排斥的问题，利用产生 Dolly 羊的核移植技术可以解决此问题。卵母细胞的核将被最终移植受体的核代替，并从形成的囊胚中分离 ES 细胞（图 21.2）。理论上讲，这些细胞可用于产生干细胞、组织甚至整个移植的器官，

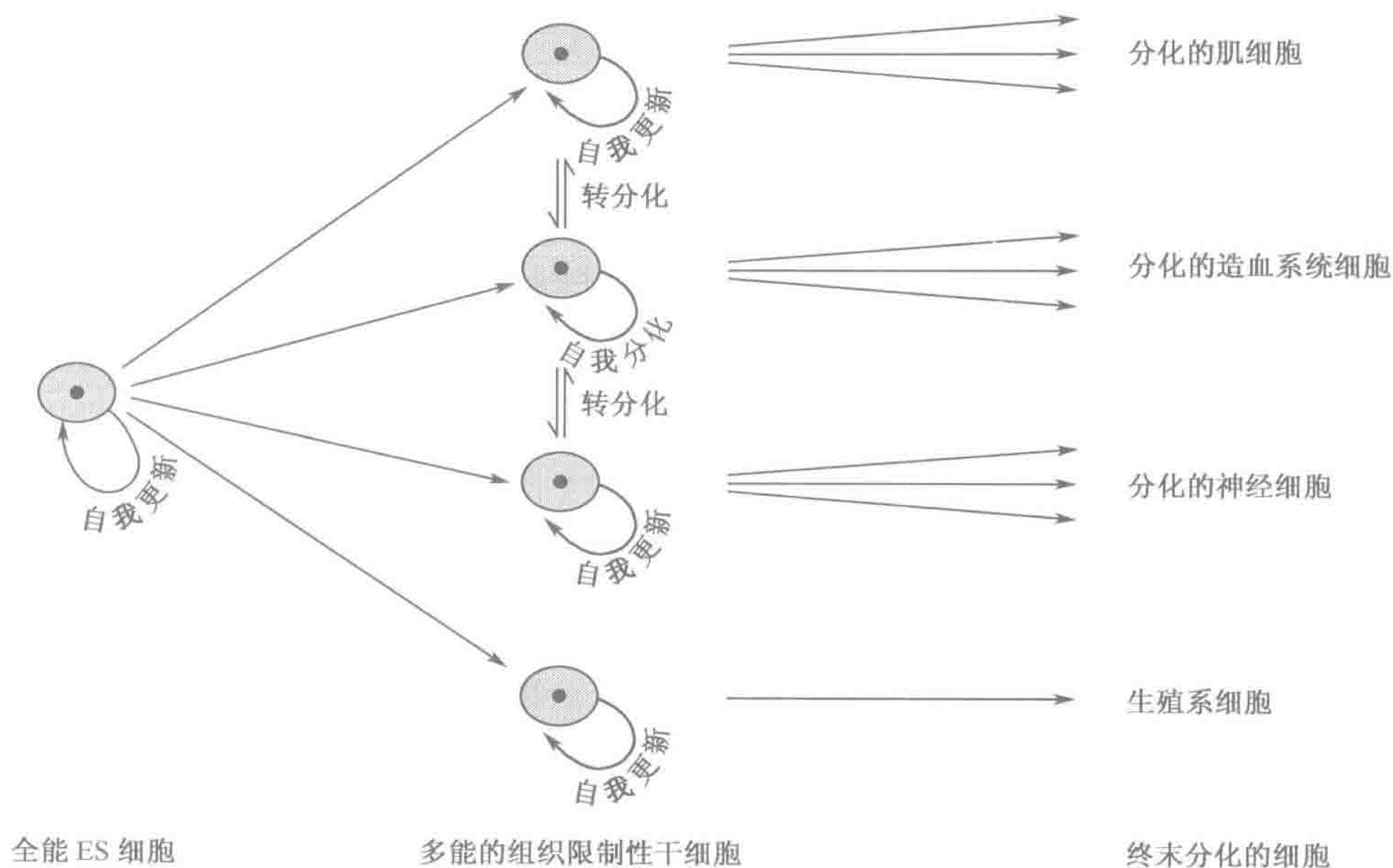


图 21.1　干细胞

所有的干细胞都能够自我更新并产生更多分化的后代。胚胎干细胞（ES）能产生生物体中所有的体细胞和生殖系细胞类型；组织限制性的干细胞分化潜能有限。组织限制性干细胞是否能够转分化成不同组织的特异细胞以及达到何种程度，目前还存在争议。

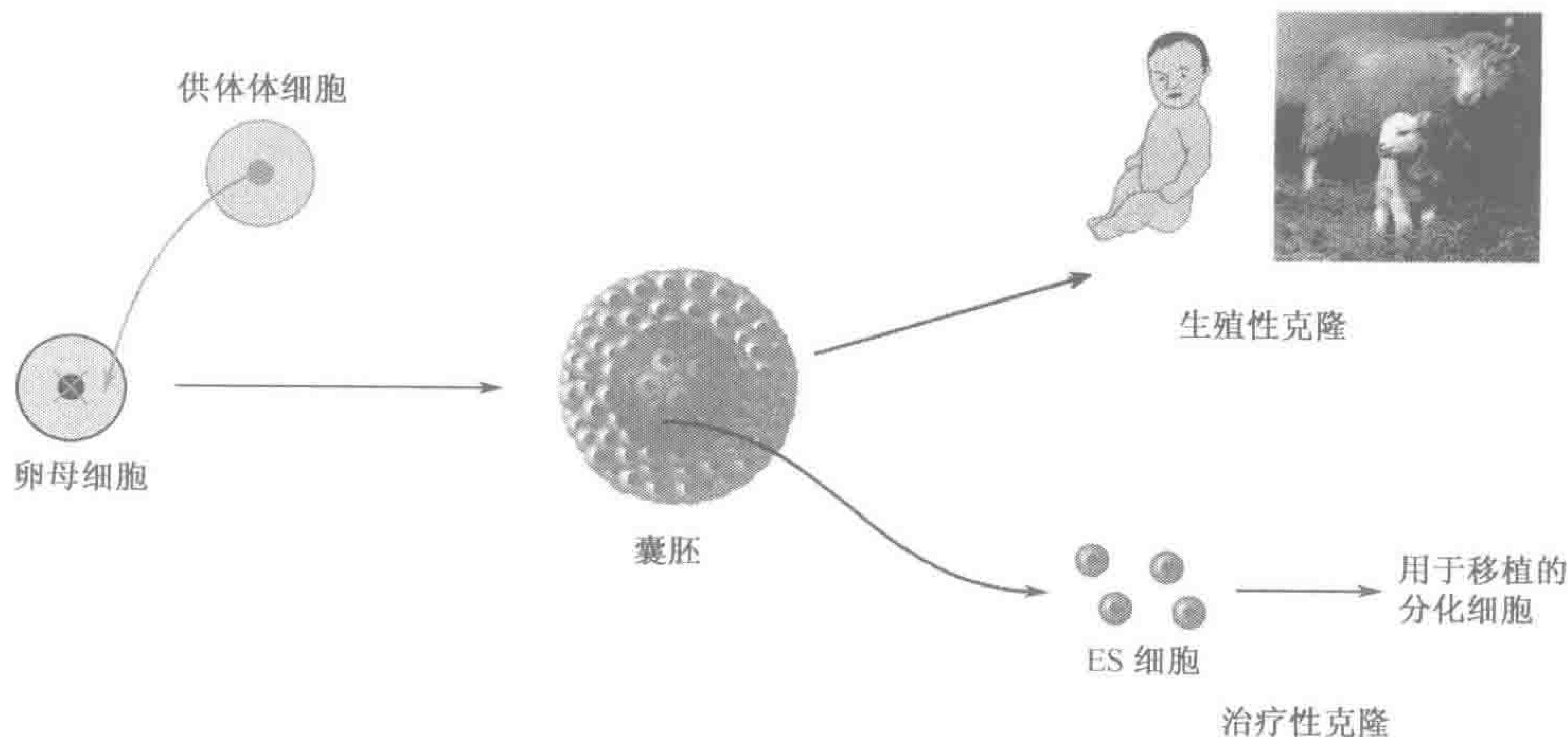


图 21.2　生殖性克隆和治疗性克隆

在这两个方法中，供体的体细胞核被植入到去核的卵母细胞，然后刺激其发育到囊胚阶段。对于生殖性克隆，囊胚被植入子宫，希望其发育成新生儿。人类生殖性克隆在大多数国家是被禁止的。对于治疗性克隆，从囊胚的内细胞团获取 ES 细胞，囊胚因此被破坏。ES 细胞作为克隆供体细胞的来源，也有可能是克隆供体组织或器官的来源。人类治疗性克隆还存在争议，因为它涉及一个人类胚胎的破坏。胚胎可以是临床体外受精（IVF）的剩余胚胎，也可能为了某种目的特别创造的。

而没有排斥的风险。这一方法称为治疗性克隆，是激烈的伦理学和政治的辩论的主题(见伦理框 1)。

伦理学框 1　人类克隆的伦理学

关于人类克隆的伦理学争论需要区分生殖性克隆和治疗性克隆——尽管最终的决定是两种类型都是违反伦理的。

许多国家的法律已考虑议案，或者试图禁止各种类型的克隆，或者试图区分生殖克隆（禁止的）和治疗克隆（允许的）。

生殖性克隆（reproductive cloning）　目的是产生一个克隆婴儿。反对生殖性克隆基于三点：实践的，原则性的，误解的。

▶**实践的争论**　指向所有哺乳动物克隆试验的低成功率以及许多出生的克隆动物有严重的畸形的事实，可能是由于目前的方法无法可靠的重现供体细胞的表观遗传修饰（Dean *et al.*，2001)。这是个简单的事实，显然以目前的知识水平任何生殖性克隆的尝试将会是严重违反伦理的。可以想像随着知识的提高可能会推翻这种反对，尽管并不清楚没有进行违反伦理的试验下我们能够如何了解。

▶ **原则的争论**　认为人类应该尊重自己的价值，而不应该将人作为达到某种目的的工具。我们完全接受该争论，但很想知道为什么它不等效地适用于那些决定他们 3 岁的孩子将成为乒乓球冠军或小提琴艺术家的雄心勃勃的父母。

▶ **误解的争论**　认为克隆不是整个个体或者不是整个人。关于生殖性克隆的许多忧虑和幽默是建立在把克隆作为某种可调鸡尾酒的基础上。这很奇怪，因为我们都知道而且非常清楚克隆是完整的人和个体。设计的克隆的一致性甚至还不如单卵双生子，因为供体很可能是一个年长的自恋亿万富翁，追求获得虚假的永生，然而克隆可能要年轻 60 岁，并进入一个完全不同的环境。举个平常的例子，即使今天有人成功的克隆了希特勒，也没有理由想像这个克隆人将着手灭绝犹太人。

支持生殖性克隆的争论是基于生殖性克隆是一对夫妻拥有孩子的唯一选择的背景——例如，如果一位妇女患有严重的线粒体疾病，她或其伴侣的一个体细胞的细胞核移植到捐赠的卵母细胞中（将提供线粒体)，那么她就可以怀孕了。

治疗性克隆（therapeutic cloning）　（图 21.2）涉及核来自于供体体细胞的胚胎干细胞的产生。目的是为供体提供完全匹配的可移植的细胞、组织或器官来源。这里的伦理学争论更加复杂。一些人反对特别为了此目的而创造胚胎的想法，甚至反对利用临床体外受精的剩余胚胎。其他反对者引用了“滑坡”理论：如果允许治疗性克隆的话，那么就不可能阻止不审慎的操作者将克隆胚胎转变成生殖性克隆项目。与此相反，支持者争辩说如果治疗性克隆是治愈致死性疾病的唯一途径，不继续进行就是不符合伦理的。

21.3.4　重组蛋白和疫苗

重组蛋白可通过对微生物和转基因家畜的表达克隆生产

一些以前来源于动物和人的治疗性蛋白可作为重组蛋白通过表达克隆生产（Russeell and Clarke，1999)。表达克隆的技术及问题在节 5.6 描述。1982 年重组人胰岛素

首先投入市场，表 21.2 列出了随后的一些例子。利用重组蛋白可以避免许多自然抽提产物的安全问题。一些血友病患者由于使用 HIV 污染的人Ⅷ因子而得了艾滋病，而一些儿童在注射了从死尸的垂体腺中提取的生长激素后死于（克雅 Creutzfeld-Jakob）病。

细菌是表达克隆最简单的宿主。能够进行大规模培养并且可以限定和控制培养基以避免污染。然而细菌产生的多肽通常不同于天然的人类蛋白质。大多数治疗性的人类蛋白质是糖基化的，而细菌不能复制人的糖基化模式。因此人们对哺乳动物表达系统的兴趣提高了。这可以是细胞培养，但转基因动物是个诱人的选择。例如一个已克隆的人类基因可以和绵羊、山羊或猪的产乳蛋白基因融合，插入到动物生殖细胞系。这就带来“药物学”构想。可以饲养乳中含有高水平期望的人类蛋白的转基因动物群（Velander *et al.*，1997）。另外，也能制造按照需要产蛋丰富的转基因鸡。

转基因植物也变得流行。植物不能复制人特异性的糖基化模式，但是它们在花费和安全性方面有许多利于表达克隆的优势。例如生产为了防御维生素 A 缺乏而设计改良的水稻品系具有相当大的利益，而维生素 A 缺乏是非洲、亚洲和拉丁美洲中至少 26 个国家的一个严重公共健康问题（Ye *et al.*，2000）。

表 21.2　通过表达克隆获得的药学产物例子

产物	用于治疗
胰岛素	糖尿病
生长激素	生长激素缺陷
凝血因子Ⅷ	血友病 A
凝血因子Ⅸ	血友病 B
α 干扰素	毛细胞白血病；慢性肝炎
β 干扰素	多发性硬化症
γ 干扰素	慢性肉芽肿疾病患者的感染
葡萄糖脑苷酯酶	Gaucher 病
组织纤溶酶原激活剂	血栓病
粒细胞-巨噬细胞刺激素	化疗后中性粒细胞减少症
来普汀	肥胖
促红细胞生成素	贫血

遗传工程抗体

B 淋巴细胞复杂的基因重排（节 10.6）使得我们每个人具有大量的不同抗体的所有组成部分，作为防御系统对抗不计其数的外来抗原。抗体分子以适配器发挥功能：在可变的（V）末端它们有外源抗原的结合部位，在恒定的末端（C）有效应分子的结合部位。抗体的结合本身可能足以中和某些毒素和病毒，但更常见的是结合抗体后可以触发补体系统和细胞介导的杀伤。

人工制造的治疗性抗体设计为单一特异性的，通常这些单克隆抗体（mAb）由杂

交瘤分泌，而杂交瘤由免疫的小鼠或大鼠产生抗体的 B 淋巴细胞与小鼠永生的 B 淋巴细胞肿瘤细胞融合产生的永生细胞。杂交瘤以独立的克隆进行繁殖，每一个克隆都能提供永久稳定的单个 mAb 来源。不幸的是这种方法产生的 mAb 的治疗潜能有限。尽管能够制造啮齿动物的 mAb 以对抗人的病原体和细胞，但是它们在人血清中的半衰期短并且可以诱发人抗啮齿动物抗体的产生。此外，不同类型抗体中只有一部分可以触发人的效应器功能。

这些问题可以通过**抗体工程**（antibody engineering）来解决。免疫球蛋白基因的不同外显子编码抗体分子的不同结构域，因此在 DNA 水平的外显子混编以构建抗体新的蛋白质组合结构域。抗体工程的早期应用是产生**人性化抗体**（humanized antibody）（图 21.3），该啮齿动物 mAb 或多或少的由人类的等同部分替代。最终构建的抗体只含有啮齿动物 mAb 的**互补决定区**（complementarity determining region，CDR）——即抗原结合部位的超变序列（图 21.3B）。最近的发展通过应用噬菌体展示技术回避了杂交瘤的构建（Hoogenboom *et al.*，1998；节 5.6.2）。这些系统允许用于构建的抗体结构域的革新性组合。一类非常有前途的抗体衍生物是单链可变片段（scFv；图 21.3C）。它们具有一个 mAb 几乎所有的结合特异性，除了可以在细菌、酵母甚至植物细胞中大规模生产单个的非糖基化多肽（Stöger *et al.*，2000）外。Hudson（1999）综述了目前开始进入临床试验阶段的许多有前途的抗体工程相关的分子。

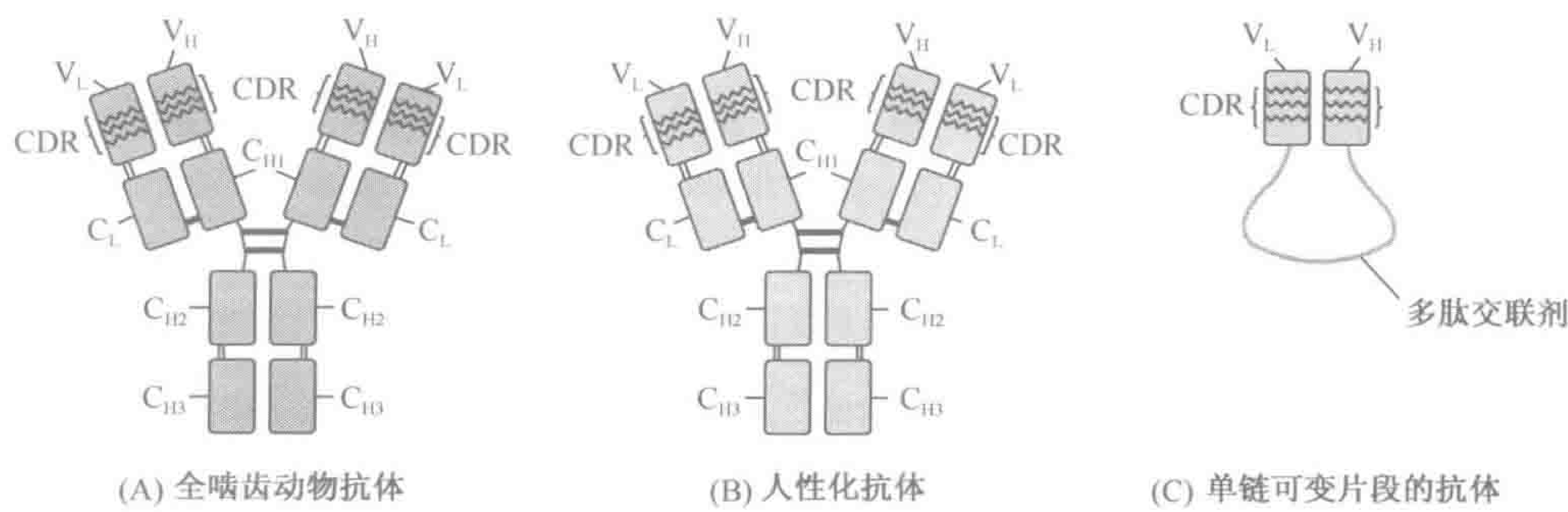


图 21.3　抗体工程

（A）经典的单克隆抗体（mAb）是单一特异性的啮齿类抗体，由来源于免疫小鼠或大鼠的 B 细胞和永生的小鼠 B 淋巴细胞肿瘤的细胞融合而成的杂交瘤合成。人类可以产生中和 mAb 的抗-啮齿类抗体。（B）这一问题可以通过工程化的 mAb 来解决，以至于除了决定特异性的超变互补决定区（CDRS）外所有的分子都来自于人。（C）scFv 是设计的单个多肽链。依赖于交联剂的长度，它们可与靶标结合形成单体、二聚体或三聚体。多聚体与其靶标结合强于单体。

针对不同背景下模仿抗体的结合特异性的相关进展。工程抗体基因可用于生产非分泌性的可与细胞内特异分子结合的**细胞内抗体**（intrabody）（Marasco，1997）。如果一个简单的结合作用可抑制靶标，则细胞内抗体是有用的。这可能是一个非常有用的特点，因为细胞内抗体不像其他小分子，不限定作用于蛋白质的特定类型（激酶、离子通道等）。此外，当抗原发生结合时细胞内抗体可以用来携带激活特定功能的效应器分子。最佳的例子可能是 caspase 3 和细胞内抗体的融合（Tse and Rabbitts，2000）。制作 caspase 3 结合的细胞内抗体用来对抗疾病相关的融合蛋白的每一半独立的蛋白部分，

诸如 bcr-abl。在癌细胞中，两个细胞内抗体都与融合蛋白结合，这使得 caspase 3 二聚化，因此触发带有融合蛋白的细胞选择性凋亡。

重组抗体的替代是利用作为特定的靶蛋白结合伴侣的短多肽（Hoppe-Seyler and Butz，2000）。实际上，在支持 DNA 或 RNA 分子（**适体，**aptamer）时，可完全抛弃蛋白质-蛋白质的相互作用。扩增 DNA 及大规模生产所需分子的能力，是利用 DNA 或 RNA 的主要优势。从一个大的随机寡核苷酸池开始，选择和扩增的重复循环用来分离具有与常规抗体相似亲和力的结合、抑制靶蛋白的分子（White *et al.*，2000）。

遗传工程疫苗

DNA 重组技术应用于抗原和抗体。病原微生物可以通过遗传修饰失活可使减毒活疫苗安全使用。遗传修饰的植物可以用来生产可食用的疫苗。改变抗原以提高免疫系统对它们的识别，结果产生一增强的应答。许多 DNA 或 RNA 疫苗目前处于临床试验，它们通常设计为肌肉内注射后高水平表达病原体或肿瘤抗原的代表性质粒（Reyes-Sandoval and Ertl，2001）。

本节对遗传知识和技术在医学及药学研究中非常广泛的应用做了扼要的综述，我们将更详尽的考虑在疾病过程中通过基因治疗进行直接的遗传干预的前景。

21.4 基因治疗的原则

为了达到治疗的目的，基因治疗涉及对患者细胞直接的遗传修饰。修饰细胞的类型和有效修饰的类型有基本的区别。

▶ **生殖系基因治疗** 产生永久的可传递的修饰。可能通过对配子、合子或早期胚胎的修饰来实现。生殖系基因治疗在某些国家由于伦理原因而被禁止（见伦理学框 2）。

伦理学框 2 生殖系和体细胞基因治疗

生殖系基因治疗涉及制造能世代传递的遗传改变。这最可能通过植入前胚胎的遗传操作来实现，但是也可能发生以体细胞为目标，附带也影响患者生殖细胞的治疗副产品。体细胞治疗仅处理患者某些体细胞，而对生殖系没有任何影响。由于单纯技术原因，生殖系治疗目前不是一个理想选择。但是，当技术问题解决后仍存在伦理学争论。生殖系的遗传操作在许多国家是被法律禁止的。

▶ 支持生殖系治疗的理由是它可以彻底地解决问题。如果能同样很好地消除风险，为什么将患病的风险遗留给患者的后代呢？

▶ 反对生殖系基因治疗的理由是这些治疗必然是实验性的。我们不能预见每一个结果，而且通过确保其作用只限于我们所治疗的患者并危险被降低到最小。这将意味着一旦我们掌握了足够的体细胞治疗的经验，进行生殖系治疗会是符合伦理的。然而，虽然最初的治疗有望是知情同意的，但后代却无法选择。这会造成一个观点，即我们有义务不把自己的思想和产品强加给后代，结果总是按此论点的话，生殖系治疗将总是违反伦理的。

支持的理由需要与群体遗传背景相比较。节 4.5 部分陈述的因素在这里高度相关；此外还有一个强有力的实际理由，即生殖系治疗是不必要的。

伦理学框 2　生殖系和体细胞基因治疗（续）

- 对于隐性疾病，仅很小部分的致病基因由受累者携带；大部分存在于健康的杂合子。哈迪-温伯格（Hardy-Weinberg）平衡给出了群体中携带者和患者的比率为 2pq∶q^2，其中 q 是致病基因频率，而 p=1−q。因为每一个携带者具有一个拷贝的致病等位基因，而每个受累者有两个；在受累者中存在的致病基因比例为 $2q^2/(2pq+2q^2)$，简化为 q。因此对发病率为 1/10 000 的隐性疾病（$q^2=1/10000$，q=0.01），只有 1%的致病等位基因是在受累者中。我们是否阻止受累者传递其致病基因（或是通过生殖系治疗，或是因治疗价格而做出绝育的不成熟选择。）对后代的发病率几乎没有影响。
- 对于完全外显的显性疾病，所有的致病基因由受累者携带，而对于 X-连锁隐性疾病的比例是 1/3。但是从群体中彻底地消除这种疾病的梦想破灭了，因为在框 4.7 的公式表明大多数严重的显性或 X-连锁疾病在群体中主要由回复突变维持。
- 第三，一个有说服力的异议是生殖系治疗是不必要的。候选夫妻很可能有显性或隐性的孟德尔疾病（再发风险分别为 50%和 25%）。假定一瓶皿含有 6 个来自此夫妇的 IVF 胚胎，选择受累的胚胎并使它们经历一个不确定的过程，而不是简单的选择 50%或 75%未受累胚胎用于重新植人，这似乎很疯狂。

体细胞治疗比生殖系治疗风险小的论点似是不可辩驳的。特别是更安全的非整合性载体不可能用于生殖系治疗。对于将我们的选择强加给后代是不合伦理的普遍观点，我们是不太认同的。可能是这样，而实际上我们一直这么做。如果政府表示同等关注不使后代承受气候变化或大量的人口过剩，将会比较容易认真的接受此观点。

- **体细胞基因治疗**　目的是以限于某个患者的方式修饰其特定的细胞或组织。目前所有的基因治疗的试验和方案都是关于体细胞治疗的。

体细胞的修饰可以通过若干不同的方式（图 21.4）。

- **基因添加**　（也称基因增加）目的是提供某个缺陷基因的一个有功能的拷贝。可以

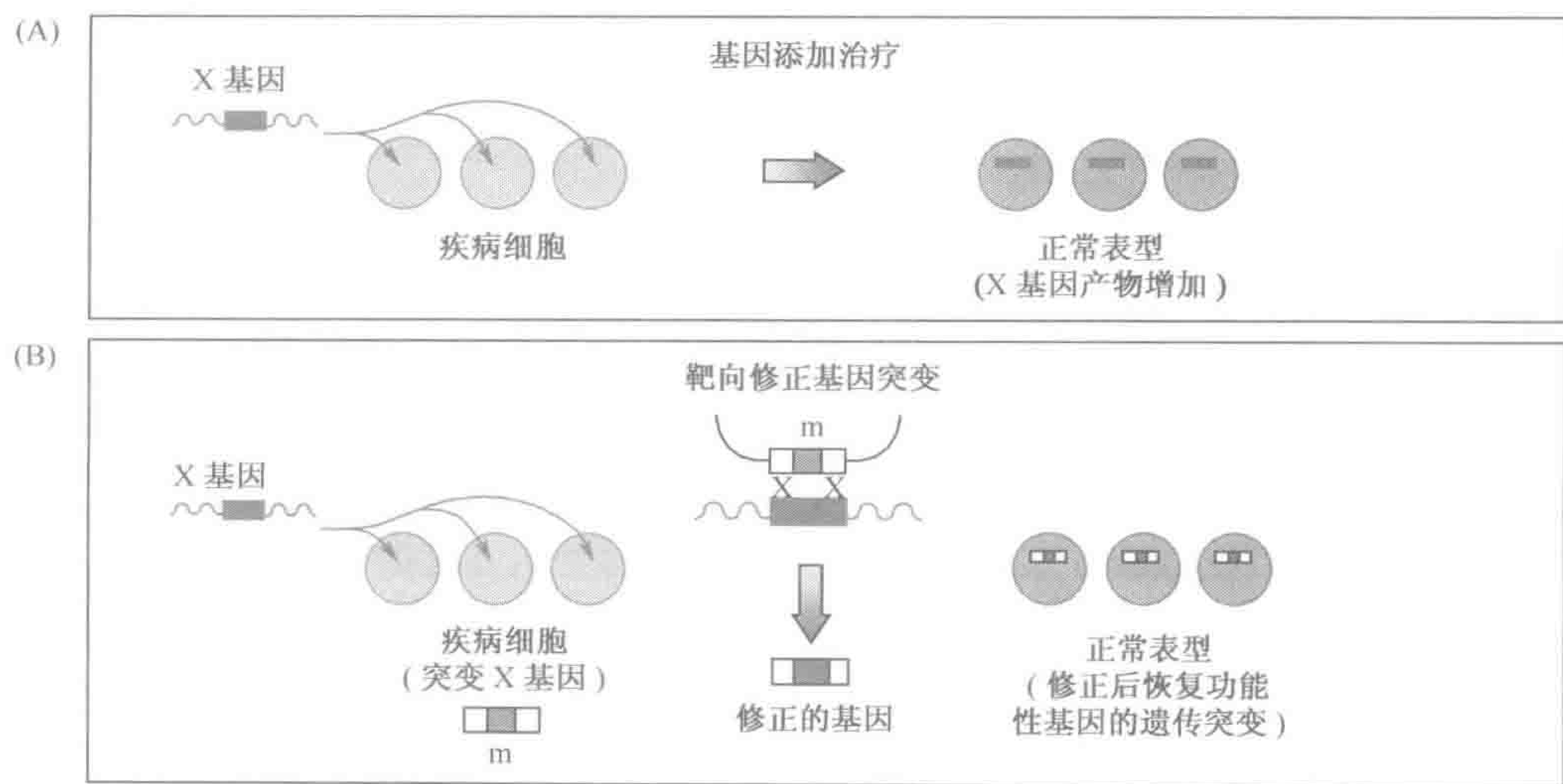
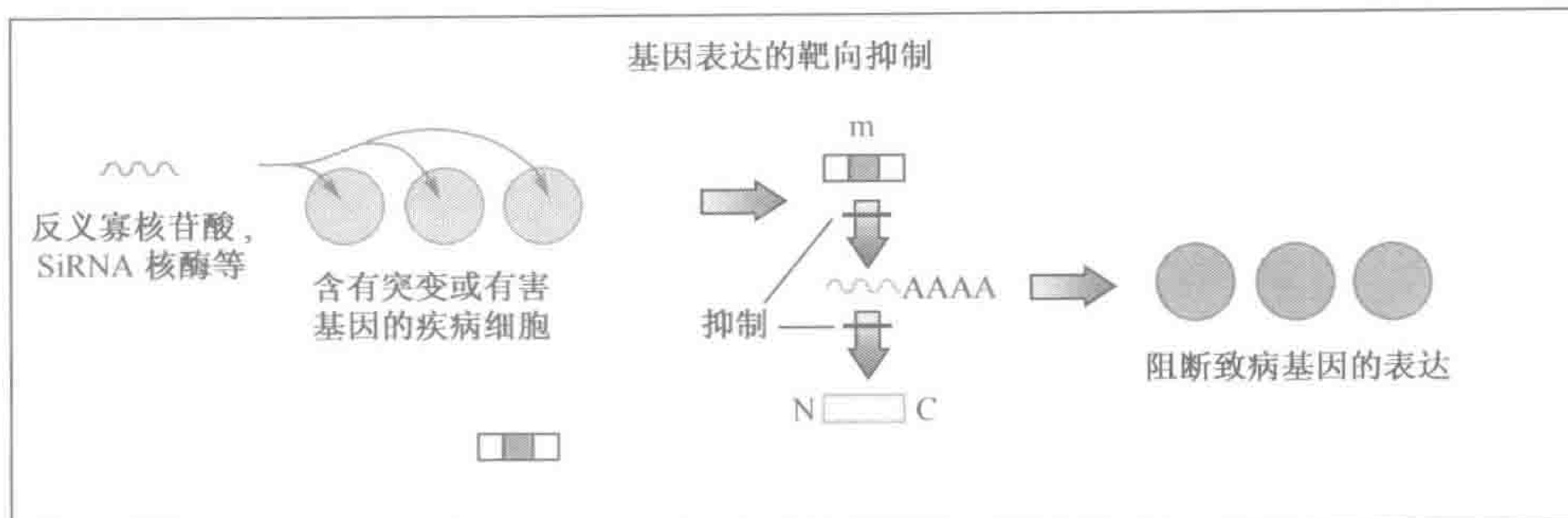


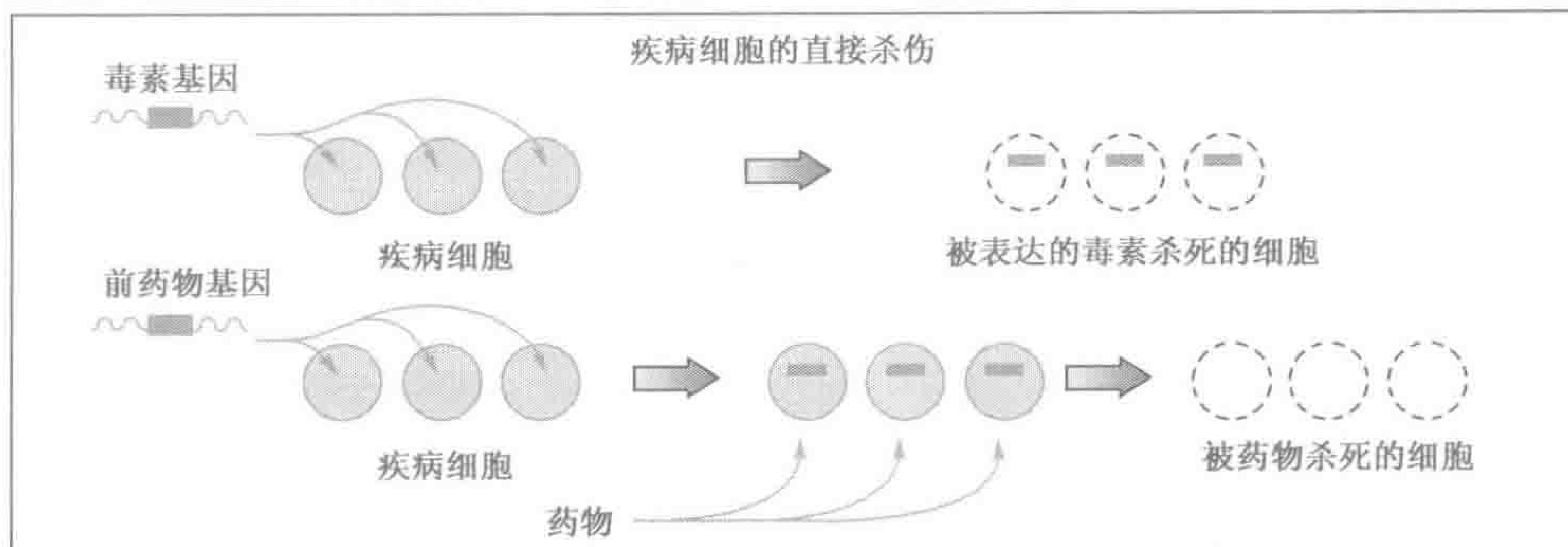
图 21.4　基因治疗的策略

详见正文。SiRNA，小干扰 RNA，见节 21.6。

(C)



(D)



(E)

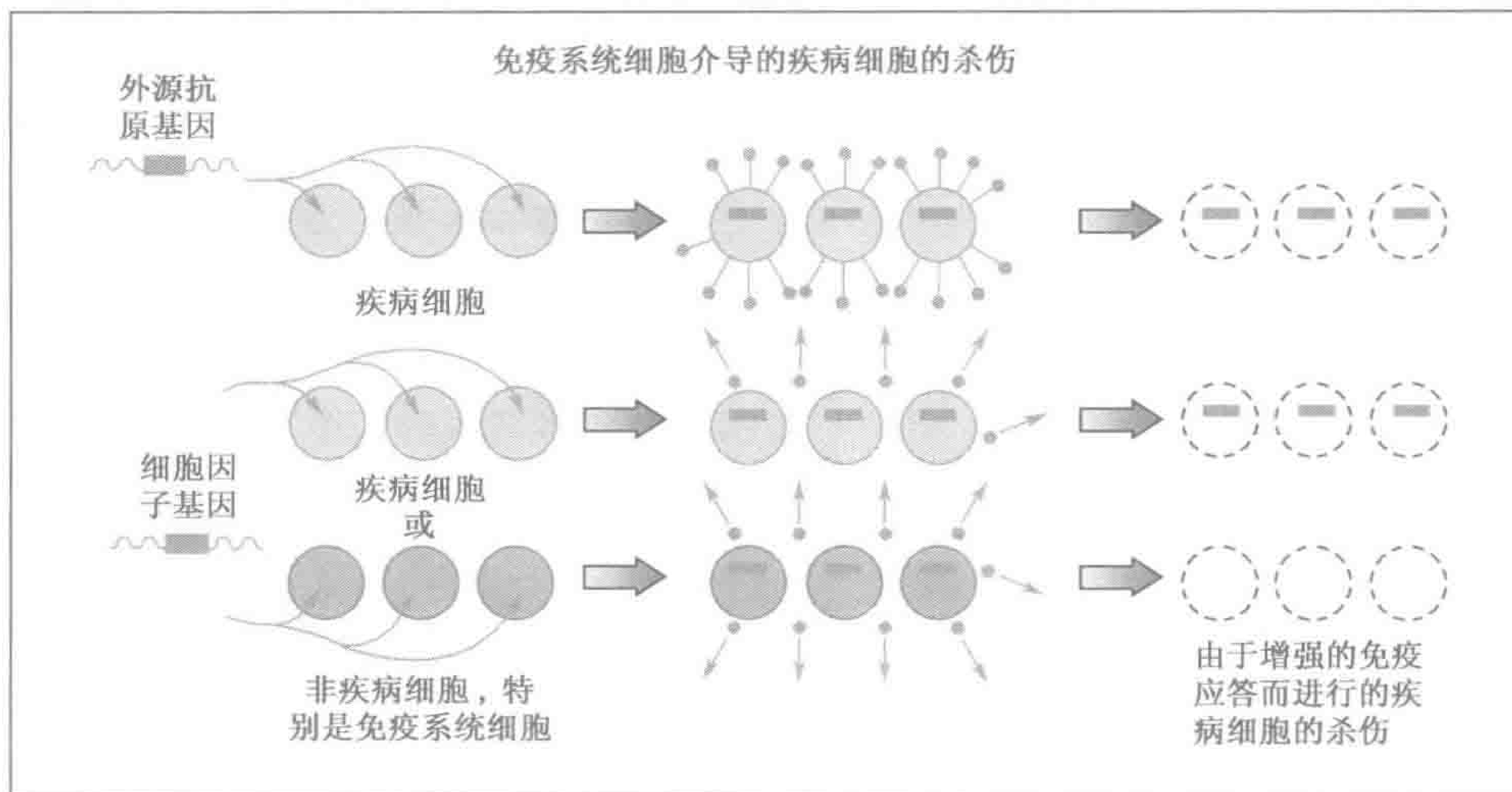


图 21.4 （续）

用来治疗功能丢失性疾病（节 16.4），此时疾病的发生是基因无功能的结果。囊性纤维化是典型的候选疾病。基因添加不适于已经发生不可逆损伤时的功能丢失性疾病，例如胚胎发育阶段的一些缺陷。癌症治疗可能涉及基因添加以提高对肿瘤的免疫应答或代替有缺陷的肿瘤抑制基因。

- **基因替代**更加有效：目的是用一个正确的有功能的拷贝来代替一个突变基因，或者原位改正一突变。基因替代适用于功能获得性疾病，这时存在的突变基因会带来一些确定坏的影响。
- **基因表达的靶向抑制**尤其和感染性疾病有关，病原体的基本功能是作用的目标。基

因表达的靶向抑制也可用于沉默癌症中活化的癌基因，消除自身免疫疾病中不想要的应答，以及在遗传病中沉默功能获得性突变等位基因。

▶ **特定细胞的靶向杀伤**特别适于癌症的治疗。

由于过度乐观继而过渡悲观的循环，对基因治疗的期望在过去的 15 年已经经历了一个躁狂——抑郁的过程。美国国立卫生研究院 1995 年的一篇重要报道试图引入某些现实（框 21.1）。从那时起，我们才看到了乐观的前景，仅在 Jesse Gelsinger 死亡时受到冲击，而成功地治疗了严重的联合免疫缺陷儿童后又重拾信心，然后由于有第一、第二例儿童几乎确切地发生了治疗后的白血病的报道而再次停滞不前（这些事件详见下文）。可能这些夸大反应的一个原因是对这类工作的自然时间表的困惑。由于诊断性检测通常能在一个基因克隆的几周内开始，人们可能会认为基因治疗也不遥远了，然而实际上这需要几十年的时期的药物发展。

框 21.1　1995 年 NIH 专门小组关于基因治疗的报道（Orkin-Motulsky 报道）

由 NIH 召集一个专门小组，以评估基因治疗目前的状态和前景，而且对未来此领域 NIH 倡议的研究提出建议。它报道于 1995 年 12 月（http://www4.od.nih.gov/oba/rac/panelrep.htm）。评估后结果有：

▶ 体细胞基因治疗是基础生物科学应用于医学的一个合乎逻辑和自然规律的进步，并且长期提供额外的管理和纠正人类疾病的潜力，这些疾病包括遗传的和获得性疾病、癌症和 AIDS……；

▶ 虽然基因治疗的期望和前景很好，但尽管有自诩成功治疗的轶事和启动 100 多个已批准的方案，在这一时期的任何基因治疗方案中并未明确地说明临床功效；

▶ 在基因治疗的所有基本方面仍有重大的问题。基本水平上的主要困难包括所有当前基因转移载体的缺点和不充分了解这些载体和宿主之间的生物学作用；

▶ 研究人员及其主持者对实验室和临床研究结果的过分吹嘘——他们可以是学术的、联邦的或是产业的——已造成了对基因治疗的误解和普遍的直觉，即基因治疗进一步发展比实际更成功。这样不准确的描写威胁了对整个领域的信心，并最终可能阻碍基因治疗在人类疾病的成功应用。

尽管已有一明确的治疗但今天仍旧需要更多的研究（节 21.7.1）

发展实用的基因治疗是一个长期的过程；另外，成功进展的报道引发了以“婴儿设计”这一构想为中心的伦理学关注（见伦理学框 3）。然而，一些学术性和商业性实验室正在致力于这方面的研究，已经通过了 600 多份关于基因治疗的实验方案。图 21.5 显示了在http://www.wiley.co.uk/genetherapy/clinical 的统计学资料。Templeton 和 Lasic 的书（进一步阅读）更为深刻地涵盖了本章剩余部分所涉及的许多问题。

伦理学框 3　婴儿设计

“婴儿设计”这一吸引人的话包含了两方面的担忧

▶ 人们可以利用体外受精和植入前诊断来选择具有特定期望质量的胚胎并淘汰即使是正常的剩余部分。这与利用避免有严重疾病的婴儿出生相同方法形成对比。

▶ 人们可以利用本章描述的治疗技术，不是为了治疗疾病而是为了遗传增强，例如赋予正常人遗传上超常的特性。

伦理学框 3　婴儿设计（续）

第一个情节已经通过植入前的性别选择和一些众所周知的病例的形式伴随我们，在这些病例中一对夫妻为了挽救有病的孩子而试图确保他们下一个孩子能够提供一个精确匹配的移植物。现在的病例涉及来源于脐带血的干细胞移植。那对婴儿没有伤害——如果打算摘取一个肾脏，那将会不同了。通常有人认为这些病例是一个滑坡的开始，即不可避免地会导致对婴儿基因型非常广泛的需求——就像短语“婴儿设计”中所设想一样，这是错误的。简单地选择 HLA 相容性就意味着四个胚胎中只有一个被选择。大多数 IVF 技术仅产生少数的胚胎，通常移植 2～3 个以确保最大的成功率。基于多条标准的选择不是简单地与具有足够的用于移植的胚胎相一致。更常见的是自然赋予我们一种简单并且高度一致制造孩子的方法，这对大多数夫妻是非常有效。而很难想像，大多数人们会放弃这个方法而支持一个耗时的、不愉快的、高侵入性的要花费大量财富但成功率低的方法。

遗传增强是一个难题——或者一旦我们对哪些基因要增强有了任何想法，它将成为一个难题。一方面，父母们应该竭尽所能为孩子提供良好的生命开端；而另一方面，由于购买一种不公平利益，随处可见仅对富人可得的选择权，至少在英国如此。或许幸运的是，即使能得到使用它们的技术，我们距离发现合适的基因很遥远。尝试生产遗传增强的动物还未成功，而有时则是惊人的失败(Gordan，1999)。从长远看来，这种可能性肯定非常大，而且一定面临着非常难的伦理问题。人们还没有在现实的思考这个问题的标志是他们总是把智力放在其一系列期望特征的首位——这些人从未比较过教授和足球运动员的生活方式吗？

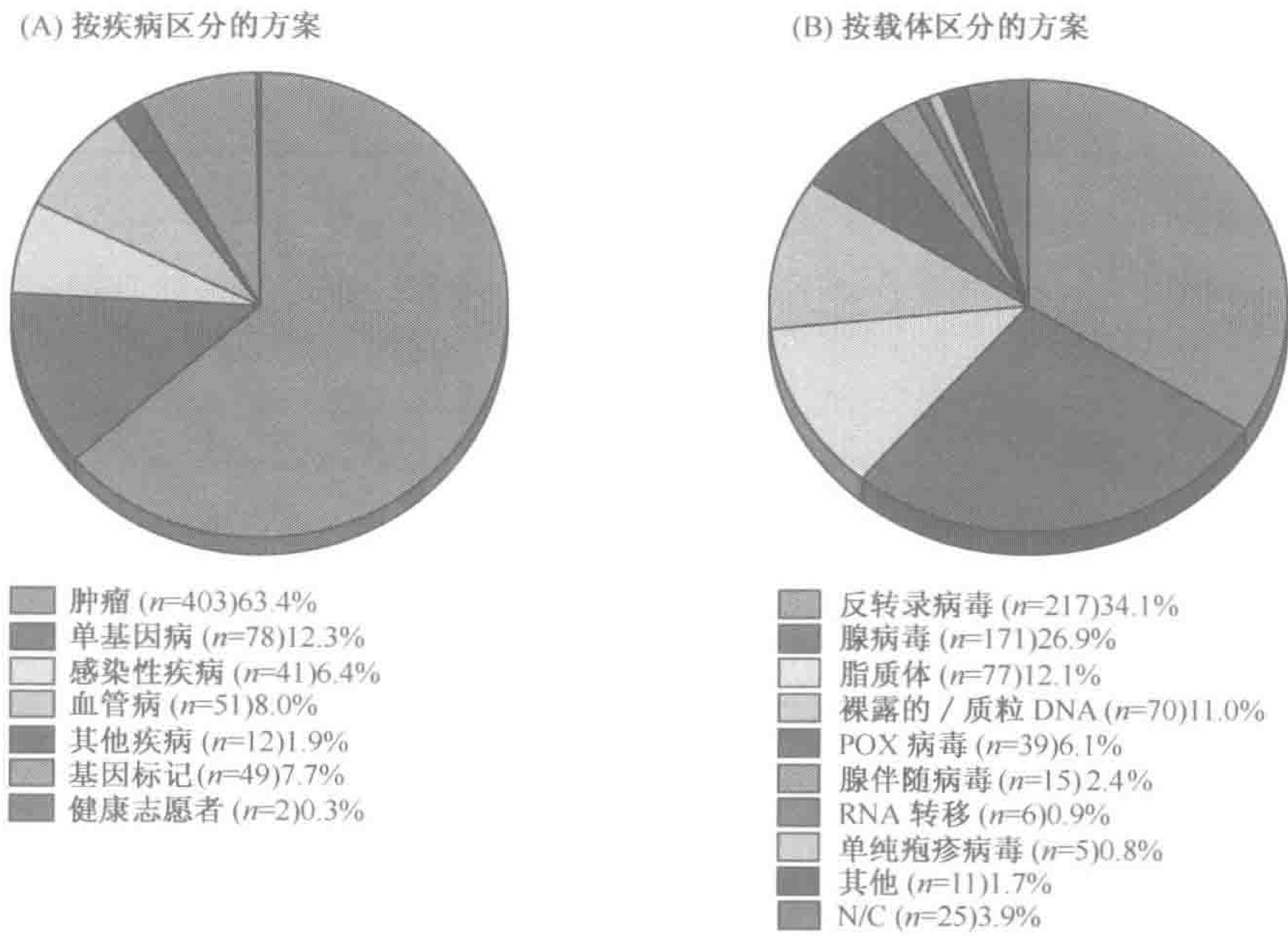


图 21.5　基因治疗的实验协议

（A）按疾病分布；（B）按载体分布本图包括 2002 年 12 月列出的所有已通过完成的、正在进行的和即将进行的实验的协议。经允许引自 http://www.wiley.co.uk/genetherapy/clinical。

21.5　在靶细胞或组织中插入并表达一个基因的方法

21.5.1　基因可以在实验室（体外）或在患者体内（体内）转移到受体细胞

体外基因转移涉及克隆的基因转移到在培养基中生长的细胞中。选择那些成功转化的细胞，经细胞培养增殖，然后回输患者体内。为了避免免疫系统的排斥应尽可能应用患者自己的细胞（自体细胞）（图 21.6）。此方法用于最初分离易于获得的、能被诱导后回输的并且替代后可以长时间存活的细胞，例如包括造血系统的细胞、皮肤细胞等。

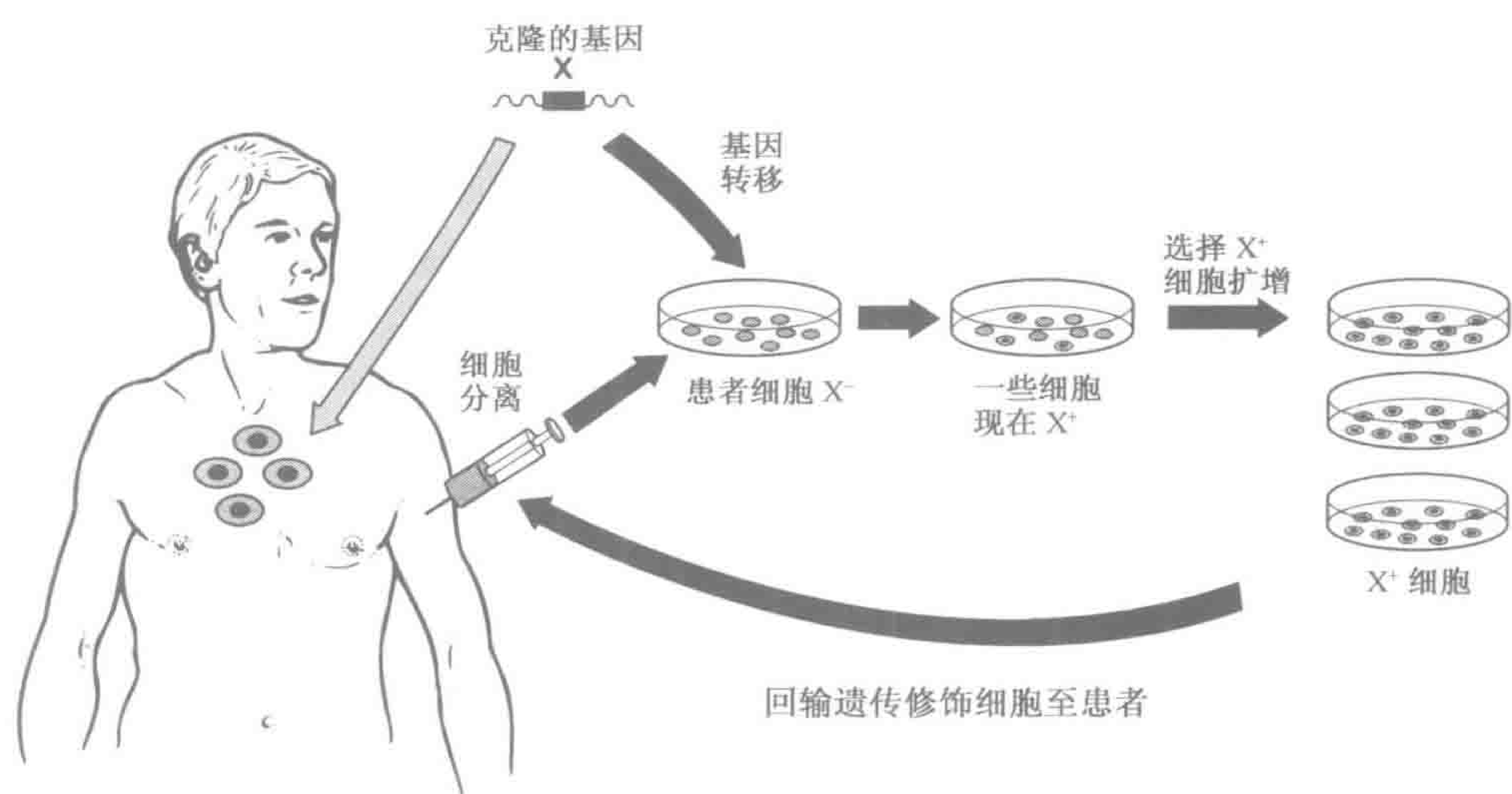


图 21.6　体内和回体基因治疗

细胞可从患者分离，在实验室中修饰后回输给患者（回体基因治疗；黑色箭头）。这仅允许合适的细胞接受治疗，而且细胞在其回输前能被检测以确保已经获得期望的改变。对许多组织来说，这是不可能的，细胞必须在患者体内进行修饰（体内基因治疗，灰色箭头）。

在受体细胞不能体外培养到足够的数目（如脑细胞）或培养的细胞不能有效地重新植入患者体内的组织中，体内基因转移是唯一的选择。靶向的组织是一个要重点考虑的事。基因转移的重组体可以直接放置到靶组织中，或者注射到全身循环中，但通过某种方式设计仅被期望的细胞类型所吸收。由于没有办法选择和扩增那些吸收和表达外源基因的细胞，因此这种方法的成功关键依赖于基因转移和表达的整体效率。

21.5.2　重组体可以设计整合到宿主细胞的染色体或作为附加体存在

为获得长期的基因表达，将外源基因整合到宿主细胞最好是干细胞的染色体上可能看来是有希望的。那么，每当宿主细胞或其子细胞分裂时，重组体就被复制。然而整合会带来某些问题和风险。大多数重组体的整合发生于随机的位置，而且在患者的不同细胞中是不同的。局部染色体环境能对重组体的表达产生难以预料的影响——重组体可能从不表达、以不希望的低水平表达，或可能短时间表达后不可逆的沉默。更糟的是整合

可能会改变内源基因的表达。插入点可能会在内源基因序列的内部，引起该基因的插入性失活。最大的担心是一个高表达的重组体的插入可能激活一个邻近的癌基因，类似于Burkitt 淋巴瘤中 *MYC* 基因的激活（图 17.4B）——实际上这恰好是在两例成功治疗了严重联合免疫缺陷的儿童身上已经发生（Check，2002，2003）。显然这两个孩子每个人的 10^6 修饰的 T 细胞中，至少有一个被反转录病毒的随机插入激活了 *LMO2* 癌基因。治疗后早期在患者的 T 细胞中有 50 个不同的插入位置，但最终这一克隆生长速度超过了所有其他的克隆，引起新型的 T 细胞白血病。如果不能证明这是这些特殊病例所用的载体或方案的某些可避免方面的结果，那么，这个经验导致随机整合载体作为工具用于基因治疗可能会全面否决。如图 21.5B 所示，这会是整个领域一个严重的倒退。

由于上述原因，以染色体外**附加体**（episome）存在的载体似乎可能成为基因治疗工具的主流。它们的缺点是基因表达持续时间有限。如果靶细胞积极地分裂，附加体将会随细胞群生长而稀释。因此不可能达到永久性治疗目的，重复治疗就很必要。为了某种目的，例如杀死癌细胞或对付急性感染，因为不需要长期的表达，所以这就不是问题。而且，如果的确出了问题，非插入的基因以某种方式自我限制，而插入到染色体的基因则无此功能。

21.5.3 病毒是基因治疗最常用的载体

没有一个基因转移体系是理想的：每一个都有其局限性和优点。然而哺乳动物病毒由于其转染人类细胞的高效性最常用于基因转移。图 21.5B 显示大约 70%已批准的使用病毒载体的方案。现已经开发了许多不同的病毒载体体系［Kay *et al*.，2001（综述)］。

肿瘤的反转录病毒载体

反转录病毒是含有反转录酶的 RNA 病毒，能够合成其基因组的一个 cDNA 拷贝。反转录病毒释放核蛋白复合体（整合前复合体）至受感染细胞的细胞质中。此复合体反转录病毒 RNA 基因组并随后将产生的 cDNA 整合到宿主细胞一条染色体的单一随机位点。整合需要反转录病毒的 cDNA 获得进入宿主染色体的通路，并只有在细胞分裂过程中核膜溶解时才能做到。因此，反转录病毒只能感染正在分裂的细胞。这就限制了潜在的靶细胞——某些重要的靶细胞，诸如成熟的神经元细胞——从不分裂。然而，只转染分裂细胞的这一特性也可以利用在癌症的治疗上。正常情况下像脑这样不分裂的组织中，分裂活跃的癌细胞能选择性的感染被杀死而对正常细胞无主要危险。

考虑到天然反转录病毒转化细胞的能力，显然设计基因治疗载体的关键是消除这种可能性。许多心思花费于设计只产生永久性失活病毒的系统。反转录病毒通常有三个转录单位，*gag*、*pol* 和 *env*，还有一个顺式作用 RNA 反应元件 ψ，该元件由包装 RNA 形成感染颗粒的病毒蛋白质识别。在载体中，*gag*，*pol* 和 *env* 被治疗基因替代，最大克隆的容量是 8kb。这个重组体在特定细胞中包装，该细胞可提供必需的 *gag*，*pol* 和 *env* 功能，但不含完整的反转录病毒基因组（图 21.7）。

反转录病毒转移 DNA 进入细胞非常有效，大多数早期基因治疗的实验都使用反转录病毒载体。然而，对插入诱变发生风险日益增高的关注使得主要的重点转向非整合的

载体上。

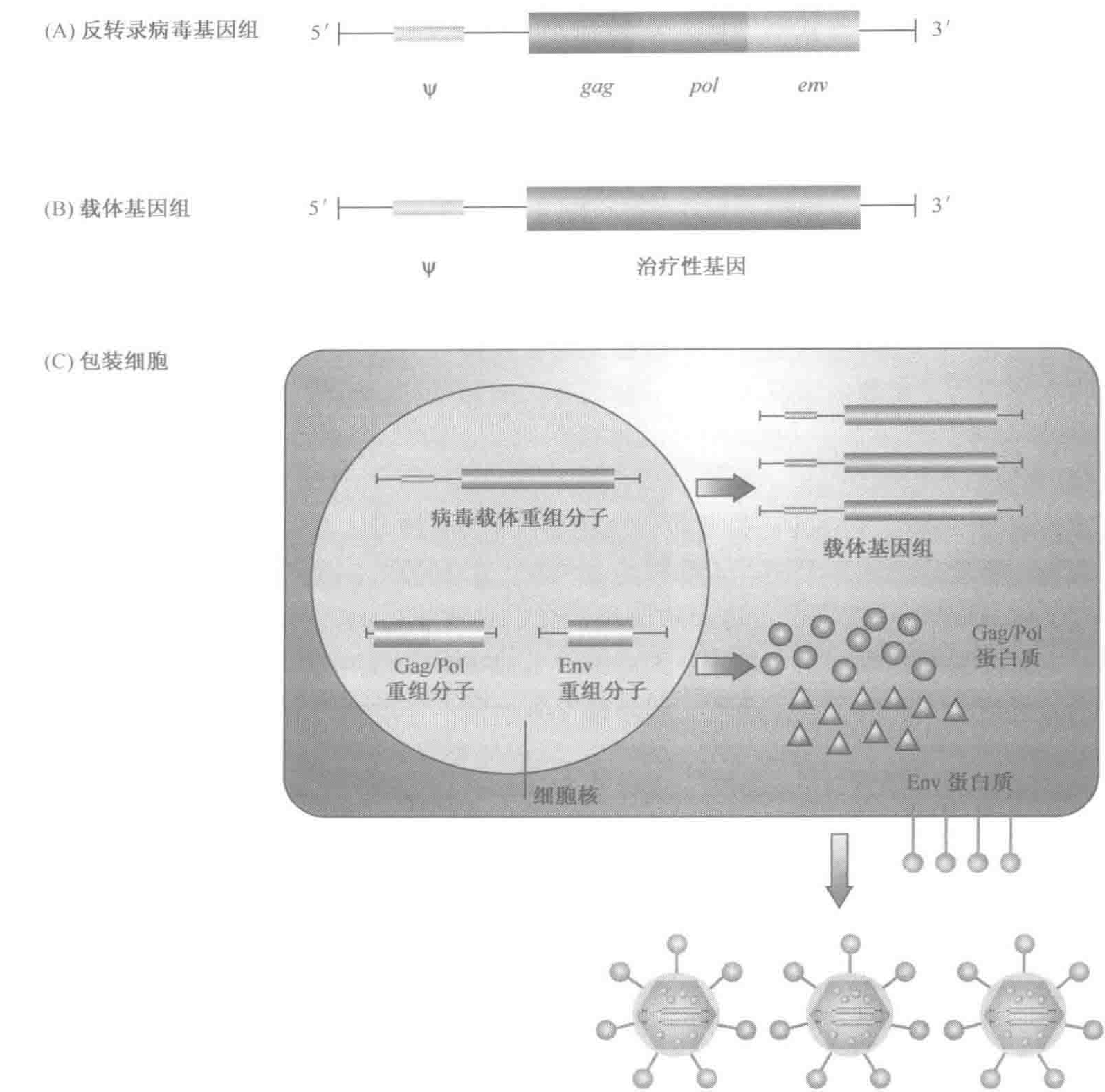


图 21.7　反转录病毒基因治疗载体的构建和包装

将 *gag*、*pol* 和 *env* 基因从反转录病毒基因组中删除，由治疗性基因替代。ψ 序列被保留，由病毒蛋白质识别，和 RNA 装配形成病毒颗粒。*gag*、*pol* 和 *env* 的功能由包裹的细胞提供，但这些基因在物理上是在独立的分子上并且无 ψ 序列。这样做是为了尽量减少产生复制有能力病毒的风险。重组病毒基因组包装成感染的但复制缺陷的病毒颗粒，病毒颗粒从细胞上出芽，从上清液中回收。重印经过 Somia 和 Verma（2000）的允许，Nature review genetics 授权，2000，Macmillan Magzine Ltd。

腺病毒载体

腺病毒是 DNA 病毒，可以引起人类上呼吸道的良性感染。它们可以大批量生产（比反转录病毒更高），有效转染正在分裂的和非分裂的细胞。线性双链 DNA 基因组在细胞核内作为附加体存在而不整合。如上所述，这在安全性上有优势，而缺点是表达持续时间短。就像反转录病毒一样，腺病毒载体是无活性的，依赖包装细胞来提供重要功能。腺病毒基因组相对较大，不同重组分子改变不同的缺失基因组部分（Yeh and Per-

ricaudet，1997）。“Gutless”载体将删除所有病毒基因，能够容纳高达 35kb 的治疗 DNA。

腺病毒载体的最大问题就是它们的免疫原性（Kafri *et al.*，1998）。尽管活的有复制能力的腺病毒疫苗几十年前就已安全用于几百万的美国军队新兵（以预防腺病毒天然感染），但在几个基因治疗的实验中有不希望的免疫反应一直是个问题。在鸟氨酸氨甲酰基转移酶缺陷症基因治疗的Ⅰ期实验中，Jesse Gelsinger 在他接受了肝内注射 6×10^{13}的重组腺病毒颗粒 2 天后，死于 1999 年 9 月。其他实验中的病人尽管损伤较轻，但有明显的感染反应。此外，由于这些载体是非整合性的，因此基因表达是短效的。在囊性纤维化的基因治疗实验中使用的第一代腺病毒载体表明 2 周后基因表达下降，4 周后表达为阴性。为维持基因表达水平有必要重复处理，但这又会加剧免疫反应。腺病毒载体将主要应用于需要基因高水平的瞬时表达，例如杀死癌细胞。

腺病毒相关病毒载体

腺病毒相关病毒（AAV）是非致病性的单链 DNA 病毒，依赖于腺病毒或疱疹辅助病毒共同感染后病毒才能复制。未修饰的人 AAV 在 19q13.3-qter 的特定位点整合至染色体 DNA。这是一个非常合乎需要的性质，提供长效表达的优势而无插入诱变发生风险这样一个腺病毒问题。不幸的是，整合特异性是由病毒的 rep 蛋白提供的，而用于基因转移的重组分子缺失了 *rep* 基因。和其他系统一样，生产病毒颗粒所必需的功能（包括 rep）是由包装细胞提供的。AAV 基础上的载体删除了 96％的 AAV 基因组。因为重组载体不含有病毒基因，因此提供高度的安全性。然而 AAV 基因组非常小，连这些高度缺失载体最多也仅能容纳 4.5kb 的插入。

慢病毒属

是专门应用于非分裂细胞的反转录病毒（Vigna and Naldini，2000）。和其他的反转录病毒一样，随机整合到宿主染色体，可供基因长效表达，但也有整合带来的所有安全问题。人类 HIV 是大多数慢病毒载体的基础，可以理解为引发非故意地产生有复制能力病毒风险的紧张。HIV 基因组比标准反转录病毒的 *gag*，*pol* 和 *env* 更复杂（图 21.13），大量的工作致力于保留感染非分裂细胞的能力的同时，减少非必需基因以及产生安全的包装系。自我失活载体提供了一个额外的安全层面（Miyoshi *et al.*，1998）。

单纯疱疹病毒载体

HSV 载体是趋中枢神经系统（CNS）病毒，这些是复杂的病毒，具有 152kb 双链 DNA 基因组，含有至少 80 个基因。能够在感觉神经节维持终生的潜伏感染，作为染色体外非整合性元件存在。这一潜伏机制用来允许转移基因的长期表达，有望通过突触网络传播。它们的主要应用将是传递基因到神经元，以治疗诸如 Parkinson’s 病和 CNS 肿瘤。插入容量至少为 30kb。实用的载体目前仍在早期开发阶段（Fink and Glorioso，1997）。

21.5.4　非病毒载体系统避免重组病毒的安全性问题，但基因转移效率普遍低

在实验室里使外源DNA进入细胞相对容易，如果病毒系统的安全问题证明是难处理的，那么某些这样的方法在基因治疗是有潜力的。然而目前所面临的是基因转移的效率低和表达持续时间短的难题。

脂质体

脂质体是当某一脂类混于水溶液时自发形成的人造囊泡。例如磷脂可以形成模拟生物膜结构的双层囊泡，亲水的磷酸盐基团在外部，疏水的脂质尾在内部。DNA在体外由脂质体包装并被转移，直接用于基因转移至体内靶组织（图21.8）。阳离子的脂质体表面带正电，在外部和DNA结合；阴离子的脂质体表面带负电，在内部和DNA结合。脂质体的包装允许DNA在体内存活，和细胞结合并被内吞到细胞内部。阳离子的脂质体是基因转移实验中最常用的载体（见Hung and Li，1997参考文献）。与病毒载体不同，DNA-脂质复合物容易制备，而且对转移DNA的大小没有限制。然而基因转移的效率低，导入的DNA也不设计整合到染色体的DNA；结果任何插入基因的表达是瞬时的。

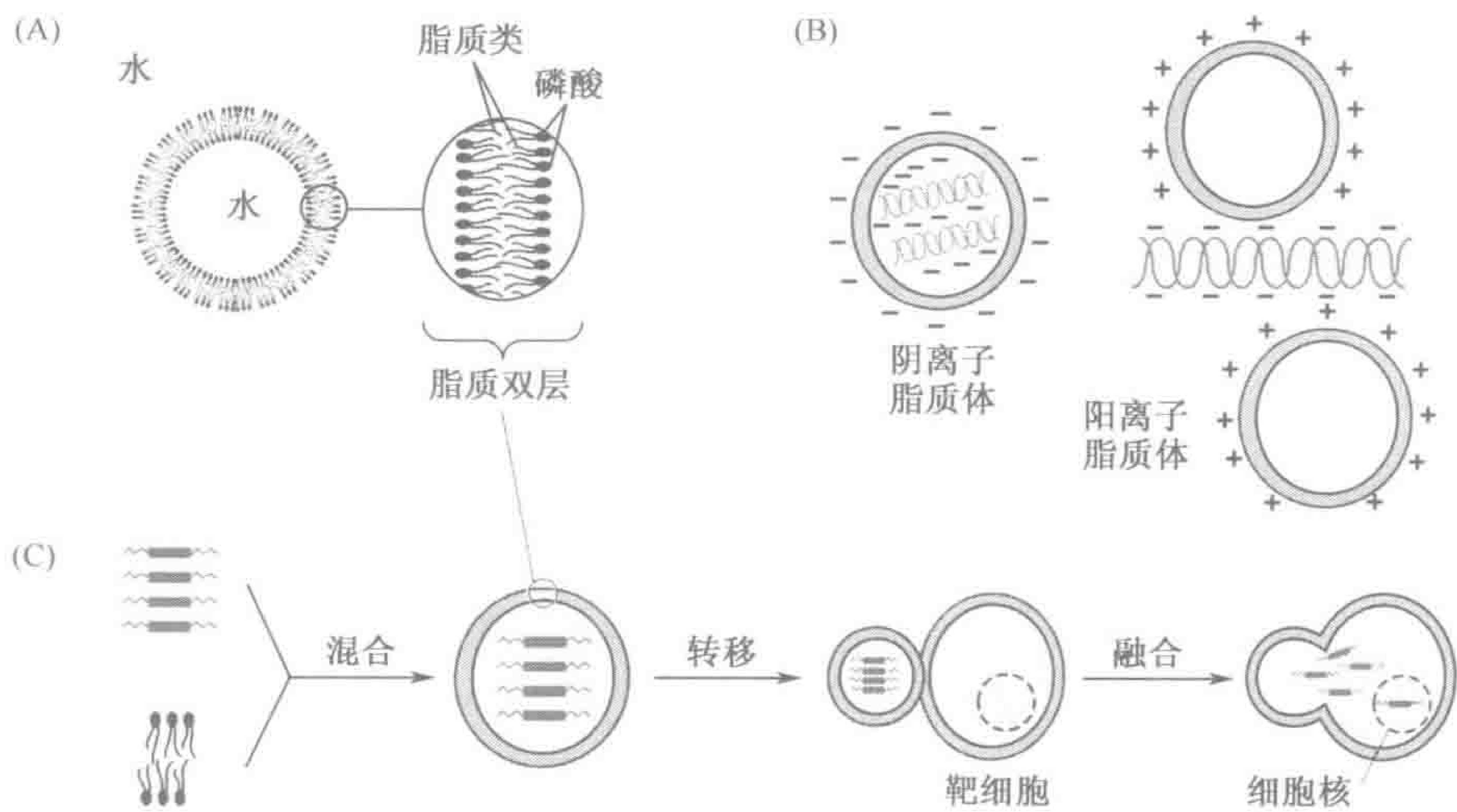


图21.8　在体内利用脂质体进行基因转移

（A）脂质体是某些脂类和水溶液混合时自发形成的人造囊泡。根据所用脂类的化学性质，它们能够携带正表面电荷或负表面电荷。（B）DNA被转运到带负电荷的脂质体的内部或结合到带正电荷的脂质体表面。（C）当脂质体和细胞质膜融合时DNA转移就发生了。

直接注射或微粒轰击

有时，DNA可以用注射器和针头直接注射到靶组织，诸如肌肉组织。例如曾考虑将这种方法用于杜兴肌营养不良。早期研究探讨了将抗肌萎缩基因肌肉注射到小鼠模型*mdx*中（Acsadi *et al.*，1991）。由于天然的抗肌萎缩基因非常大，因此使用一个小基

因——由抗肌萎缩基因的 cDNA 和确保高水平表达的调节序列，如一个强大的病毒启动子构成。另一种直接注射的替代方法就是利用微粒轰击技术（生物射弹或“基因枪”）技术：DNA 由金属颗粒包装后由特殊的枪发射到细胞中。应用此简单、相对安全的方法已经成功地将基因转移到许多不同组织中。然而利用这些直接注射方法的任何一种，注射的 DNA 不能稳定地整合，基因转染效率非常低。这在某些组织可能问题还不大，如肌肉，通常不增殖，注射的 DNA 可持续表达几个月。

受体介导内吞

在这种方法中转移的 DNA 连接到一个能够与特定细胞表面受体结合的靶分子，引起内吞将 DNA 转移到细胞内。例如肝细胞通过其细胞表面受体清除血清中的去唾液酸糖蛋白。去唾液酸糖蛋白和多聚赖氨酸共价结合后通过带正电荷的多聚赖氨和带负电荷的 DNA 之间的电荷作用，可逆地结合 DNA。如果复合体经由胆管或血管床注入到肝脏，它就会被肝细胞选择性吸收。更普遍的方法是利用转铁蛋白受体，该受体在许多细胞类型表达，但在增殖的细胞和造血细胞相对丰富（图 21.9）。

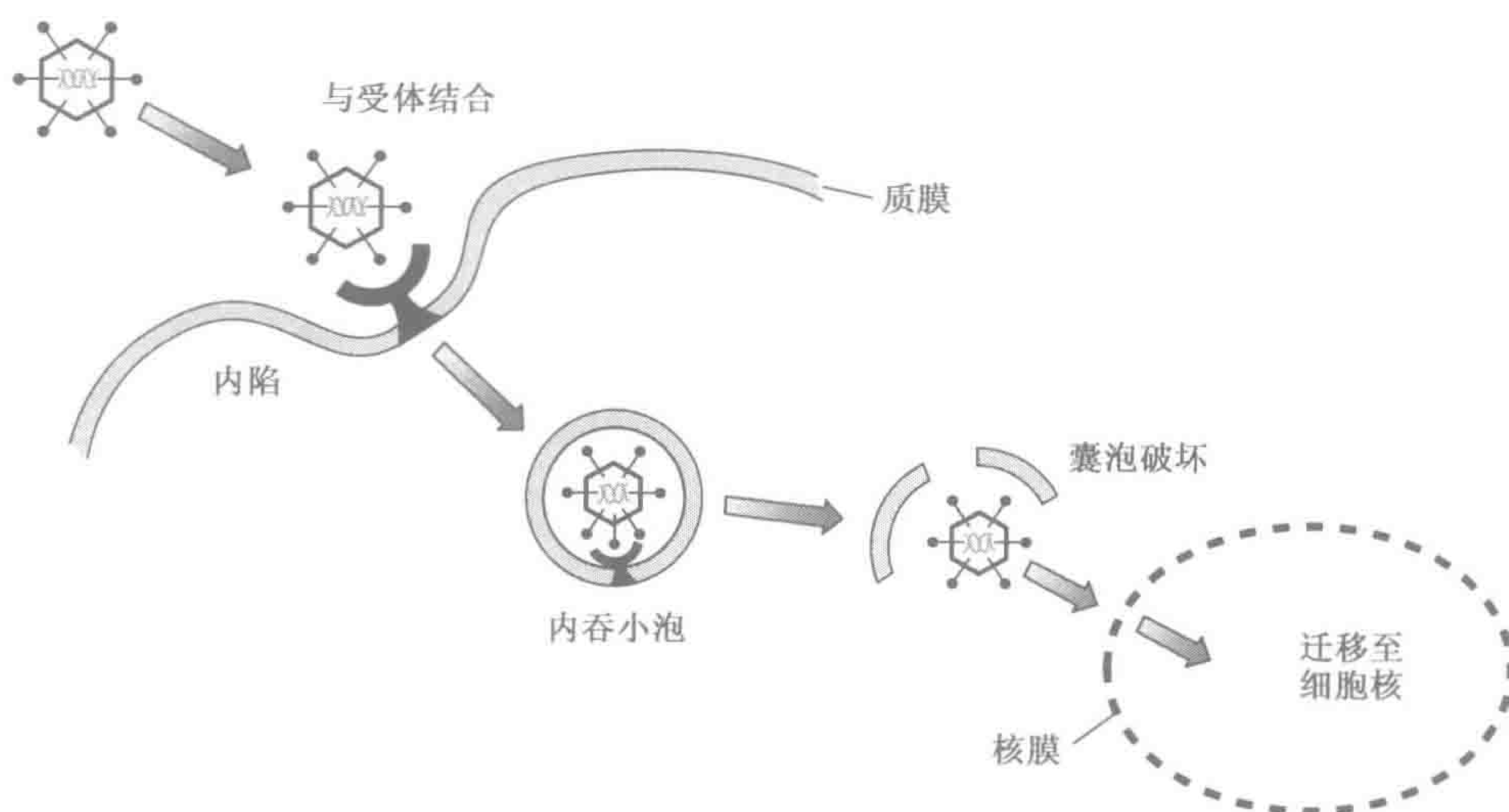


图 21.9　经由受体介导内吞进行基因转移

在和配体结合后，质膜内陷、脱落，使受体-配体复合物形成细胞内的囊泡，内吞小体。配体可以是携带治疗基因的腺病毒，如图所示；或者是治疗性 DNA 和某些其他靶细胞具有的特异性受体的其他分子结合。通常内吞小体靶向溶酶体的降解。为了表达，在此发生之前外源 DNA 必须以某种方式从内吞小体中逃避出来到达细胞核。腺病毒特异性破坏内吞小体，引起有效地逃逸。

通常以这种方式将内在化的底物传递到溶酶体，而且如果 DNA 要到达细胞核，则必须提供某种逃逸机制。一种可能就是共转移腺病毒或腺病毒蛋白。这可以特异性破坏早期的内吞小泡，使得内容物逃逸。虽然基因转移的效率可能高，但这种方法未设计转移基因的整合。

21.6 在细胞或组织中修复或失活一个致病基因的方法

上一节介绍的方法都是为了用一种或另一种法方法在靶细胞或组织中插入一个期望的基因，让它以一定水平表达并在一段时间以解决特定的临床问题。然而某些问题需要不同的方法。由功能获得的基因（节 16.5）和显性负效应引起的疾病（节 16.4.3）不会受益于这种治疗方法。如果问题是一个确实会造成某种有害影响的常驻基因，唯一的解决方法是除去或失活这有害基因。典型的例子是显性的孟德尔疾病（而非那些由单倍型剂量不足引起的疾病，节 16.4.2）、在癌细胞中激活的癌基因和在自身免疫疾病中不适当的免疫反应。感染性疾病也可能是通过抑制病原体特异性基因或其基因产物来治疗。

可以尝试各种不同的策略来达到这一目的。致病基因可能是确实地被破坏，或者通过下调转录，破坏转录物或抑制蛋白质产物阻止表达。无论应用何种方法，某种媒介物或重组分子必须进入靶细胞并且在那儿发挥功能。在那一方面，有效传递和表达的问题和前面章节所描述的相同。在显性疾病或激活的癌基因的情况下，还存在另一问题，即设计一种物质只选择性攻击突变等位基因，而不影响正常等位基因。那一领域通常被看作是基因治疗的第二个高潮。目前研究在此处以证明原理为目的：只有当基因添加治疗被认为成功后才有可能发展实践性治疗。

21.6.1 通过同源重组修复突变的等位基因

在比较低等生物中许多标准的遗传工程技术利用同源重组将一段序列代替另一段序列，因此很自然就考虑到利用同源重组来修复致病的突变等位基因。在小片段的同源重组中标准方法之一是用来传递 400～800bp 含有野生型序列的 DNA 双链到靶细胞（Goncz *et al*.，2001）。原理的证明已有报道，但尚难证实它取得足够高效率的有用治疗。

21.6.2 通过反义寡核酸抑制翻译

反义寡核苷酸，通常 12～30 核苷酸长，能够通过和互补 mRNA 形成 dsRNA 或 DNA-RNA 双螺旋抑制翻译（Galderisi *et al*.，1999）。在某些生物，诸如秀丽新小杆线虫，dsRNA 有效地转变成小的干扰 RNA 分子（siRNA），该分子通过虽不很了解但却高效的 RNAi 机制（见下文）失活同源的转录物。这在人类显然不会发生。人类细胞几乎或根本没有 siRNA 形成时所需的 Dicer 酶。相反，dsRNA 或 DNA-RNA 被 RNAse H 所破坏。这种酶降解 RNA-DNA 双链中的 RNA 链，而非 DNA 链；因此使 DNA 反义分子具有半催化的功能。反义分子通常被化学修饰使其对细胞内核酶更加有抵抗力。磷硫酰连接代替了磷酸二酯键，或在整个糖-磷酸盐骨架用抗核酸酶的合成性多肽框架代替的地方应用肽核酸。这就带来了非特异性毒性的缺点，而且反义的重组分子一定是体外化学合成的，而非修饰的反义寡核苷酸可以通过适宜的质粒在细胞内产生。反义寡核苷酸更有不可预见的效应。在某些情况中，靶基因被有效地和特异地沉默；但经常是没有什么效果或有非特异的效果。超过大约 30 bp 长的 DsRNA 常常会诱发干扰素的产

生，引起翻译的普遍停止。1998 年在 Vivatrene 用于治疗 AIDS 患者的巨细胞病毒感染的磷硫酸盐修饰的 DNA 成为首个进入市场的反义核酸药物。

21.6.3 通过核酶选择性破坏或修复 mRNA

多年来，已发现越来越多 RNA 分子具有酶功能（核酶，Doudna and Cech，2002)。例如催化性 RNA 涉及端粒酶（作用于 DNA)，内含子-外显子剪接（作用于 RNA)，以及作为核糖体中肽键转移酶（作用于多肽)。RNA 分子有许多使其适于作为酶发挥作用的特性。它们可以形成各种三维结构，依靠其能够特异地和 DNA，RNA 或蛋白质分子结合。核酶具有很多潜在的羟基和氨基基团。

基因治疗学家们对以序列特异性方式切割靶 RNA 的核酶特别感兴趣。在自然界，自我切割的 RNA 在通过滚环机制复制的 RNA 病毒的多联体的位置特异性切割产生基因组长度的链。天然的例子包括锤头（40nt)，发夹（70nt)，HDV（人类 delta 病毒，90nt）和 VS（Varkud 卫星；160nt）的核酶（Doudna and Cech，2002)。通过遗传工程可以将锤头核酶转变成多能的反切割酶，其靶 RNA 的序列可以简单地用与核酶互补的序列指定（图 21.10)。正如用反义寡核苷酸（见上文)，核酶通过化学修饰后对细胞内的核酸酶有抗性，但是它必须是外源导入的，而不能是在转染的靶细胞内通过自然转录产生。第一例核酶的 Ⅰ 期临床试验于 1998 年开始，Ⅱ 期临床试验正在应用作用于 VEGF 受体 mRNA 的核酶，抑制乳腺癌和结肠癌中血管的形成（Sullenger and Gilboa，2002)。

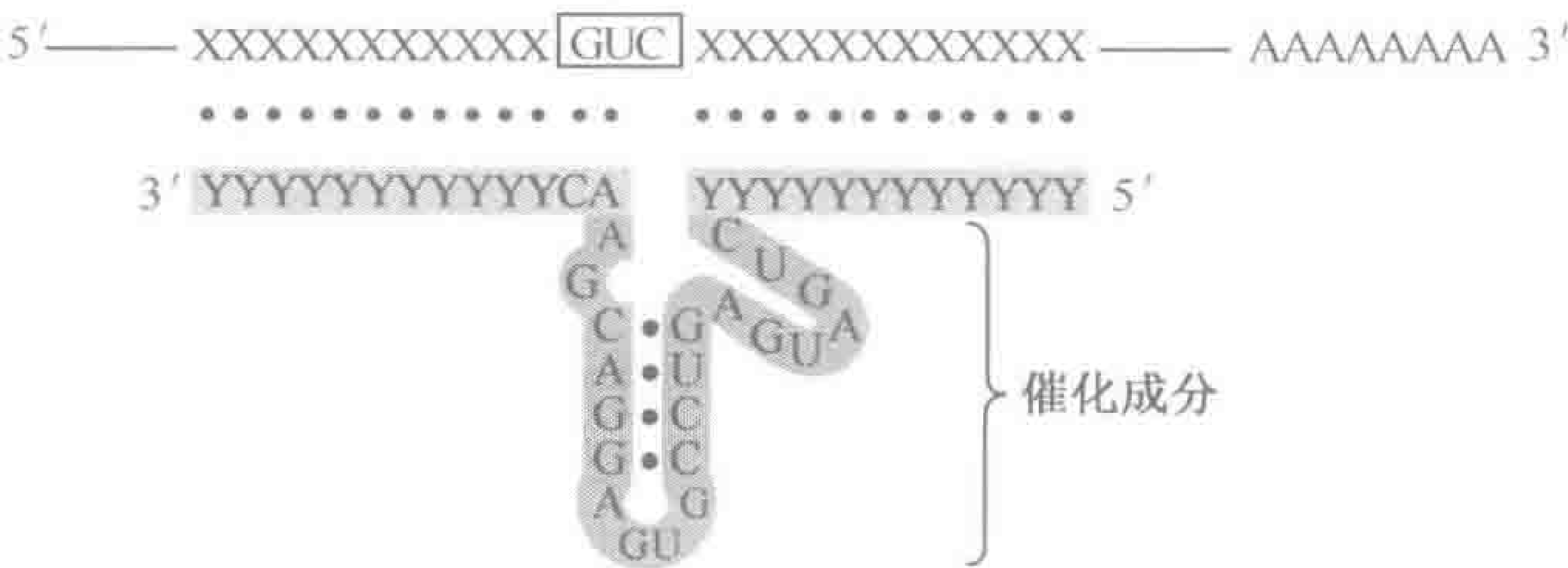


图 21.10 利用反-切割锤头核酶破坏突变的 mRNA

天然的锤头核酶是自我切割的。为了基因治疗的目的，含有切割位点的链（靶点）和催化链（人造核酶）分离。通过设计核酶的序列（YYYY……）和靶的序列（XXXX……）互补，几乎任何期望的 RNA 都能被特异性靶定。

由于核酶的序列特异性是由碱基配对来控制的，所以它们可以很容易地设计从而特异性降解显性疾病中突变的 mRNA，而保持野生型完整的转录物。更值得炫耀的是，核酶已设计通过反剪接反应来修复突变的转录物，在此反应中预先确定的突变序列从 mRNA 上被切除，并被野生型序列代替（Sullenger and Gilboa，2002)。所有这些想法已经证明在试验系统中起作用，但是否它们能够发展成为实际的治疗方法尚有待观察。

21.6.4 通过 RNA 干扰（RNAi）选择性抑制突变的等位基因

这是目前涉及最令人振奋的技术。如节 20.2.6 所描述，应用 siRNA 在许多生物可

以获得特异、高效的基因表达抑制。siRNA 为 21～23nt 的双链 RNA，5′端磷酸化，3′端有 2nt 突出。在许多生物，较长的 dsRNA 可以被 Dicer 酶有效切割成 siRNA。这在人类细胞中并不起作用，但是在人类细胞中可以由合适的 DNA 载体转录产生 siRNA。设计的转录物可以回折形成茎-环发夹。茎为 19～29bp，设计为与靶序列匹配；环为 6～9nt。目前 RNAi 是一个产生功能丢失表型的高度有效的实验工具。尽管前景令人鼓舞，但它是否也能发展成为一种实用的治疗手段还不太清楚。

21.7　人类基因治疗尝试的一些例子

图 21.5A 显示，600 个左右已获得批准的基因治疗方案中，63%是关于肿瘤的，只有 12%是关于单基因病的，另外 6%是关于感染性疾病的，8%关于血管疾病的。这里我们给出在一些重点领域中基因治疗进展的简要回顾和一些参考。尽管实验数目有限，但是单基因疾病总是在基因治疗议程主要的前列，第一个明确的成功例子就在此领域。

21.7.1　第一例明确成功的例子：治愈 X 连锁的严重联合免疫缺陷

回顾 90 年代，治疗常染色体隐性严重联合免疫缺陷（SCID）的尝试作为一个成功的例子受到公众的热烈欢呼。将携有功能 *ADA* 基因的反转录病毒载体回体转染到腺苷脱氨酶缺陷患儿的 T 淋巴细胞，在培养基中增殖后回输给患者。后来其他的患者接受了相似的治疗，每个患者经过 10～12 次治疗。报道了几例患者临床表现改善引人注目。然而，所有的患者也继续接受了酶制剂的治疗，因此不清楚他们临床症状的改善究竟多大程度是基因治疗的效果。

第一例不明确的成功来自一个相关的疾病，X-连锁 SCID（Cavazzana-Calvo *et al*.，2000；图 21.11）。治疗再一次是回体的，使用的是编码细胞因子受体基因 *IL2Rγc* 链的反转录病毒载体。表达 CD34，一个造血干细胞标志，和反转录病毒载体共同孵育 3 天，这段时间细胞数目增长 5～8 倍，然后将细胞回输给患者（图 21.11）。11 例治疗患者中有 9 例治愈，能过正常的生活（Hacein-Bey-Abini *et al*.，2002）。然而，正如上面提到的那样，他们中有两人随后发生了白血病，几乎可以肯定是 *LMO2* 癌基因的插入激活的结果（Check，2002，2003）。所有涉及反转录病毒转染大量淋巴细胞的试验在全世界范围内迅速被搁置起来，对这些悲惨事件的理解仍悬而未决。到写作本书时，还不清楚有多少会恢复。

21.7.2　囊性纤维化基因治疗的尝试

在孟德尔疾病中，囊性纤维化应该是更经得起基因治疗考验的一个。问题是由于 *CFTR* 编码的氯酸盐通道的缺乏引起的，主要位于呼吸道上皮。对有部分活性的 *CFTR* 等位基因的患者的研究，表明 5%～10%的正常水平就足以产生良好的临床反应。这至少可以阻止疾病进展，逆转某些继发性影响。组织特异性基因转移可通过使用气雾吸入器完成。至少报道了 18 例 CF 基因治疗的临床试验（Davies *et al*.，2001）。最初的试验使用腺病毒，自然情况下感染非分裂的肺细胞。高剂量时这些腺病毒有时会激发炎症反

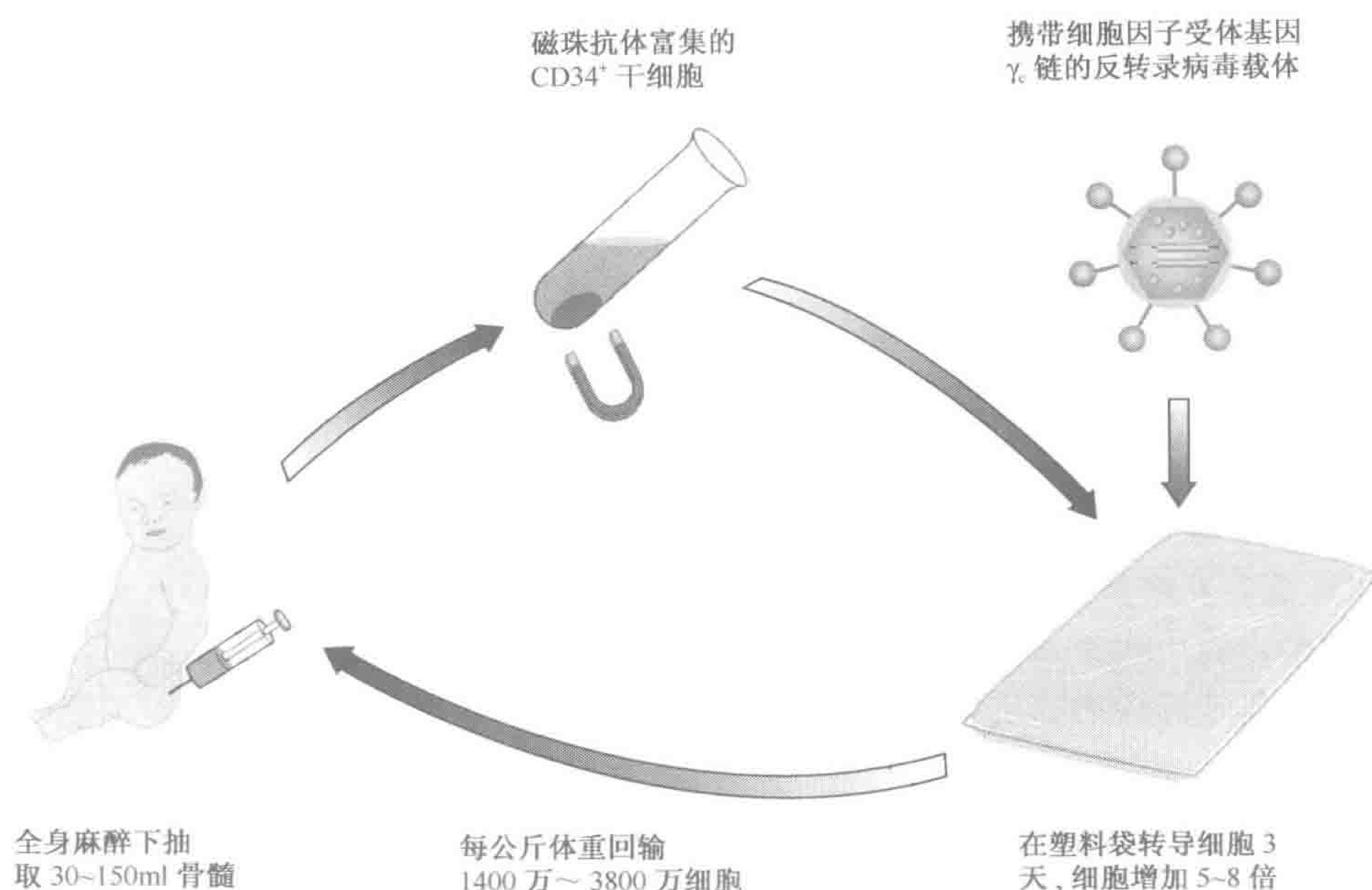


图 21.11　X 连锁的严重联合免疫缺陷疾病（X-SCID）的基因治疗

这是第一例明确的成功的基因治疗。在巴黎 Neker-Enfants Malacles 医院治疗的 11 名年龄 1～11 个月的男患儿，9 例治愈，见 Hacein-Bey-Abini *et al*.（2002）。9 例患者中有 2 例后来不幸得了白血病，几乎可以肯定是由于反转录病毒载体的临近插入激活 *LMO2* 癌基因的结果。

应，特别是腺病毒诱发 Jesse Gelsinger（见上文）死亡后，目前人们对其持怀疑态度。5 个最近的试验中有 4 个使用了脂质体或 AAV。几个试验论证了基因转移和短暂表达，但人们很清楚所有这些试验都不能很好满足临床的适用性。

一个重要问题就是覆盖在肺脏呼吸道上皮细胞的黏液和外被多糖的物理屏障，特别在 CF 患者的感染肺脏。基因治疗的制剂能传递到呼吸道，但需要更高级的运载工具以确保上皮细胞的有效转染。转染适当的细胞也是个问题。腺病毒倾向于感染上皮基底细胞，而天然的 CFER 在黏膜下腺体有高水平的表达。干细胞应该是理想的靶标，因为表面上皮细胞的寿命仅约 120 天，因此需要重复给药，并带来免疫应答所有伴随的问题。

21.7.3　杜兴肌营养不良基因治疗的尝试

对 DMD 女性携带者和轻度贝克型肌营养不良患者的研究表明恢复肌肉组织中大约 20%的正常抗肌萎缩蛋白基因表达就有利于 DMD 患者。然而，DMD 的基因治疗面临着双重问题，即抗肌萎缩基因非常大（2.4Mb），将其转移到骨骼肌和心肌细胞中都非常困难。关于进展的综述见 Chamberlain（2002）。甚至抗肌萎缩基因的 cDNA（14kb）对很多载体来说也是很大的。基于对一个具有抗肌萎缩基因 17～48 外显子（46%的编码序列）缺失，但只有轻度的贝克型肌营养不良的患者的观察，构建了一个 6.3 kb 的能容纳于腺病毒和反转录病毒载体的小基因。重要的是避免诱发细胞介导的免疫应答，

使用肌肉特异性的启动子来启动转基因有助于使这些减少到最小。腺病毒相关病毒在肌肉中表现出良好的持久性，至少在健康的小鼠中如此，质粒也是很有希望的运载工具。不仅将重组分子传递到骨骼肌，而且还要传递到心脏和膈肌；这对成功治疗很有必要。目前这一问题还没解决。另一种精致的可选择的方法是试图上调 utrophin 表达（MIM 128240），utrophin——一个类抗肌萎缩分子，正常时仅在神经肌接头处与肌肉细胞膜上肌动球蛋白复合体结合。Utrophin 由一个独立的常染色体基因座编码，此基因座在 DMD 男患儿中保持完整。

通过移植成肌细胞到患儿的肌肉中，也已尝试 DMD 的细胞治疗。在某些研究中偶尔发现的患者具有一些抗肌萎缩阳性的肌纤维，但任何一个患者都没有临床症状的改善。骨髓移植传递的干细胞可能迁移到肌肉，作为肌肉干细胞发挥作用（Ferrari *et al*., 1998）。如果得到证实，那么这就提供了利用治愈 SCID 的方法为治疗 DMD 带来真正的希望（见上文）。

21.7.4　癌症基因治疗

所有已批准的基因治疗实验方案中超过 60％是关于癌症的（图 21.5）。表 21.3 列出很多例子来说明这种方法的种类。这包括：

- 基因添加以恢复抑癌基因的功能
- 基因失活以阻止激活的癌基因表达
- 肿瘤细胞的遗传操作以触发细胞凋亡
- 肿瘤细胞修饰使之抗原性增加，以便免疫系统破坏肿瘤
- 树突细胞修饰以增加肿瘤特异性免疫反应
- 利用人工设计的溶瘤病毒选择性杀死肿瘤细胞
- 肿瘤细胞遗传修饰，以便肿瘤细胞而不是周围的非肿瘤细胞，能使无毒性的药物前体转化成有毒物质，杀死肿瘤细胞（图 21.12）

表 21.3　肿瘤基因治疗试验的例子

这些多数是一期(基本安全性)实验；在这一发展阶段，大多数试验的目的不是在临床上使患者受益。见 NIH 临床试验数据库(http://www4.od.nih.gov/oba/rac/clinicaltrial.htm)

疾病	改变的细胞	基因治疗的策略
卵巢癌	肿瘤细胞	腹膜内注射编码全长 p53 或 BRCA1cDNA 的反转录病毒或腺病毒，希望恢复细胞周期调控
卵巢癌	肿瘤细胞	注射编码 ErBb2 抗体 scFv 的腺病毒，希望失活生长信号
恶性黑色素瘤	肿瘤浸润淋巴细胞	从手术切除的肿瘤中提取 TIL，在培养基中培养增殖。用表达肿瘤坏死因子 α 的反转录病毒载体在回体感染 TIL，回输给患者。希望 TIL 会靶向作用剩余的肿瘤细胞，TNFα 杀死它们。原理见图 21.4E
各种肿瘤	肿瘤细胞	将表达细胞表面抗原，如 HLA-B7，或细胞因子，如 IL-12，IL-4，GM-CSF 或 IFNγ 的反转录病毒转染肿瘤细胞。希望增强肿瘤的免疫原性，以便宿主免疫系统破坏它们。通常通过回体应用致死性照射的肿瘤细胞实现。原理见图 21.4E

续表

疾病	改变的细胞	基因治疗的策略
前列腺癌	树突细胞	用肿瘤抗原或表达抗原的 cDNA 处理自体的树突细胞，使它们获得对肿瘤细胞的免疫应答增强。原理见图 21.4E
恶性神经胶质瘤（脑肿瘤）	肿瘤细胞	将表达胸苷激酶（TK）或胞嘧啶脱氨酶（CDA）的反转录病毒注射到肿瘤。只感染正在分裂的肿瘤细胞，而不感染周围的非分裂的脑细胞。然后用 gancyclovir（TK-阳性的细胞会将其转变成有毒的 gcv 磷酸盐）或 5-氟胞嘧啶（CDA-阳性的细胞会将其转变成 5-氟尿嘧啶）。病毒感染的细胞（正在分裂的）被选择性地杀死。见图 21.12
头颈部肿瘤	肿瘤细胞	注射 ONYX-015 人工合成的腺病毒到肿瘤。病毒仅能在 p53 缺陷细胞中复制，因此选择性地溶解肿瘤细胞。当结合系统的化疗时，这种方法非常有效

TIL，肿瘤浸润淋巴细胞；TNF，肿瘤坏死因子；IL，白介素；GM-CSF，粒细胞-巨噬细胞集落刺激因子；IFN，干扰素。

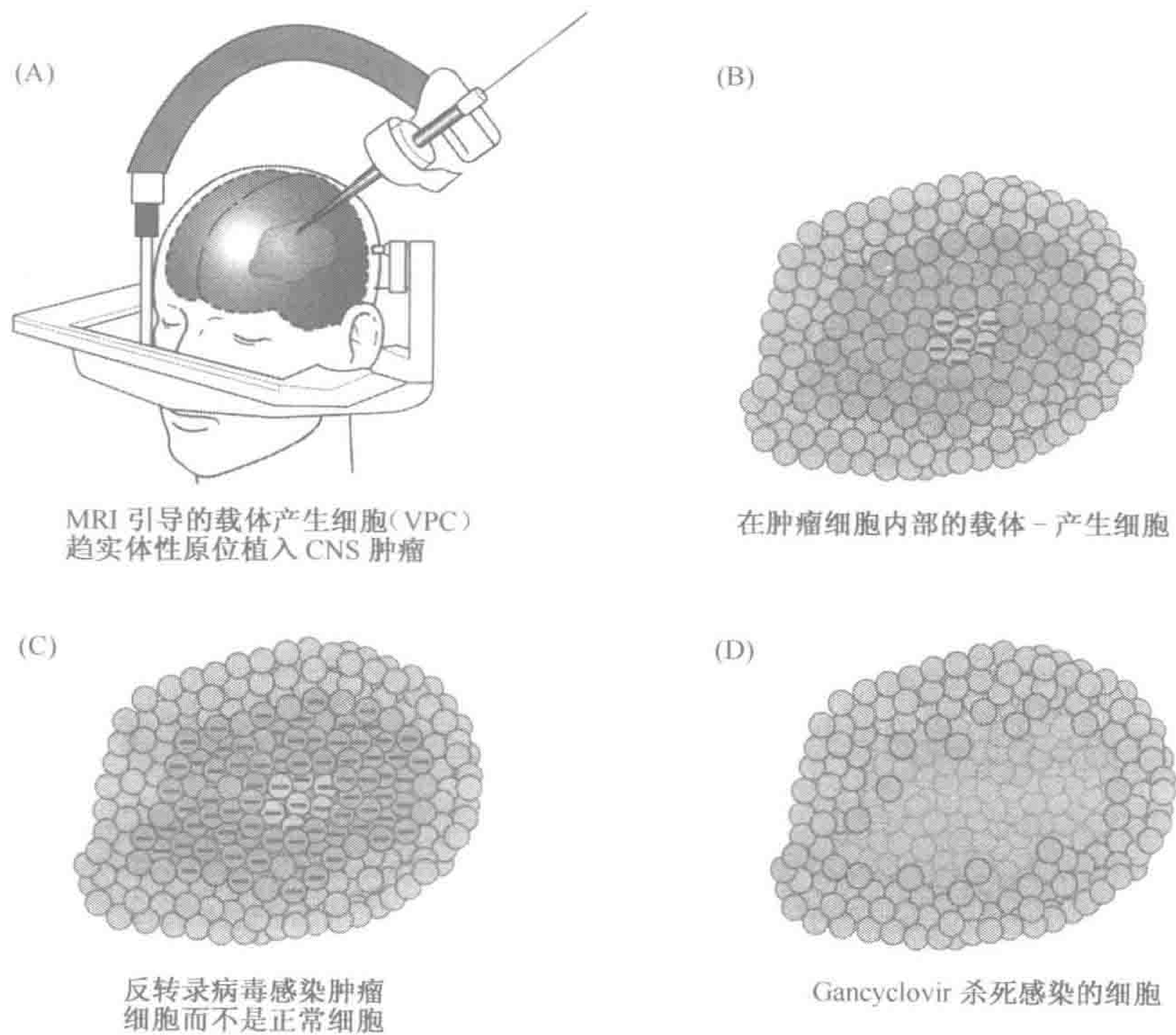


图 21.12　HIV-1 体内基因治疗脑肿瘤

设计一个反转录病毒以产生单纯疱疹病毒胸苷激酶（HSV-TK）。载体产生的细胞（VPC；蓝色）被注射到脑肿瘤中。由于反转录病毒只感染正在分裂的细胞，因此它们感染肿瘤细胞（粉色）而不感染周围正常脑组织（绿色）。血管内给予无毒性的药物前体 gancyclovir（gcv）。在 TK^{+} 的细胞 gcv 转变成高毒性的 gcv-三磷酸盐，细胞被杀死。

21.7.5 基因治疗感染性疾病：HIV

作为人类最重要的病毒病原体，HIV-1 是医学科学每一相关分支中大量研究的目标。遗传操作在两个领域很重要。大量工作投入到尝试发展遗传工程疫苗；此外许多学者认为对宿主细胞遗传操作可以使它们抵抗 HIV。NIH 的临床试验数据库（http://www4.od.nih.gov/oba/rac/clinicaltrial.htm）列出了 1992～2001 年提出的近 40 个关于基因转移的实验方案。几乎所有这些在原理上都和治愈 X-SCID 的方法相似（见上文）。将某一个很有希望抑制 HIV 复制的基因转染造血干细胞，通常使用反转录病毒载体，然后处理的细胞回输患者。由于 AIDS 的主要病理变化就是感染和淋巴细胞的破坏，因此这是一个试图阻止 HIV 感染发展成 AIDS 的一种自然方法。

HIV 是一个反转录病毒，成熟的病毒颗粒由两个相同的单链 RNA 基因组拷贝加上一些核心蛋白构成，含有包裹在由病毒糖蛋白和病毒出芽繁殖时从宿主细胞膜摄取的脂质构成的外壳下（图 21.13）。和所有反转录病毒会有 *gag*，*pol*，*env* 基因一样，HIV 复制需要 Tat 和 Rev 调节蛋白。这些，以及它们结合的病毒 RNA 序列（TAR 和 RRE）是一些使淋巴细胞抵抗 HIV 的尝试的选择。策略包括：

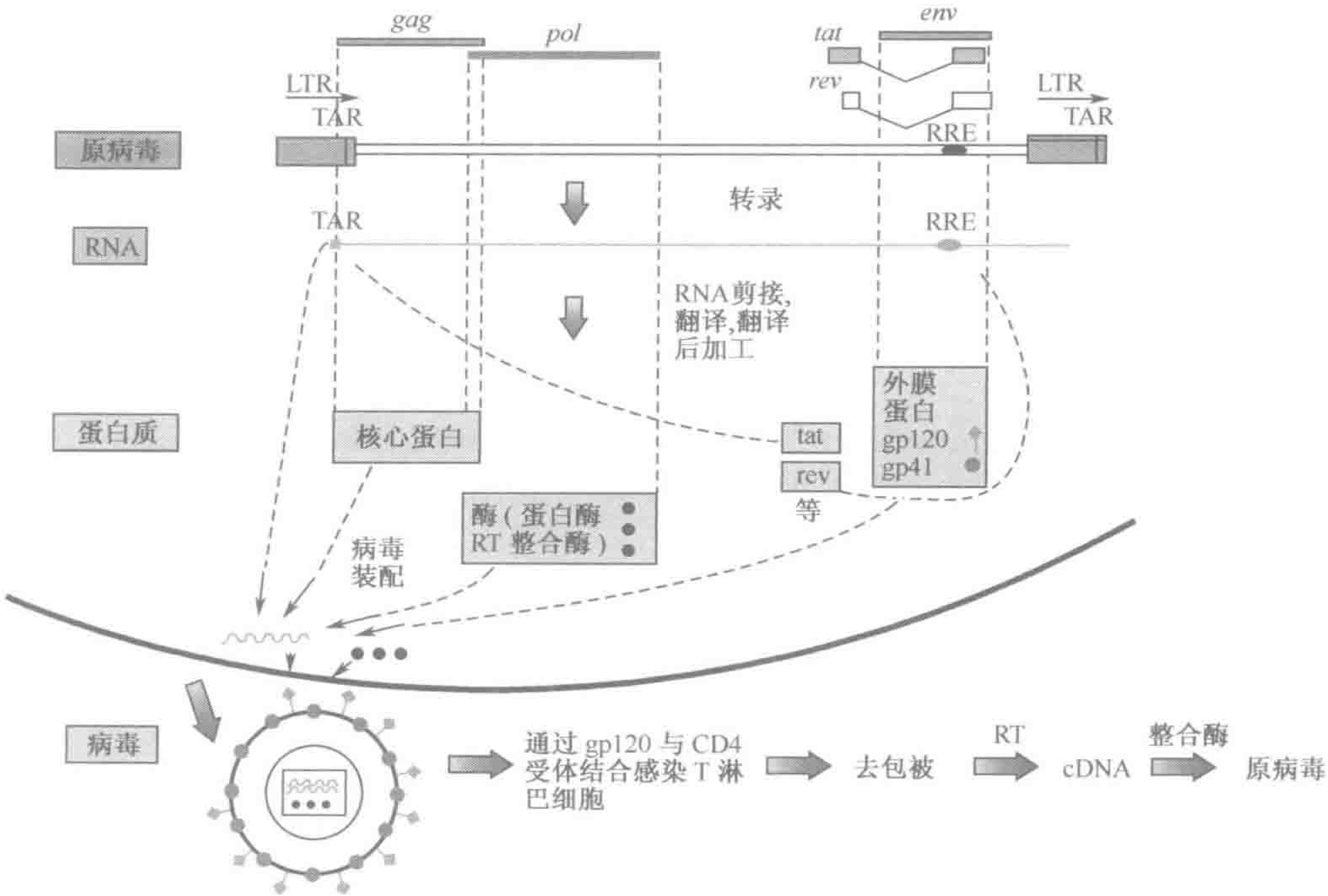


图 21.13 HIV-1 病毒的生命周期

HIV-1 是具有 RNA 基因组的反转录病毒，像其他反转录病毒一样，有 *gag*，*pol*，*env* 基因。利用反转录酶合成 DNA 拷贝后作为原病毒整合至宿主的染色体。病毒 mRNA 和蛋白质包装成病毒颗粒，从宿主细胞膜出芽分离。和简单的反转录病毒不同，HIV-1 编码两个调节蛋白，Tat 和 Rev；分别和病毒基因组的 TAR 和 RRE 序列结合，是病毒复制所必需的。这些是 HIV 基因治疗的主要靶标。

- 反义 RNA 的利用。构建编码针对 TAR，重叠的 *tat* 和 *rev* mRNA，以及 *pol* 和 *env* mRNA 的反义 RNA 反转录病毒。
- 诱饵 RNA 的利用。直接高水平表达含有 RRE 序列的转录物的反转录病毒可能能够隔绝 Rev 蛋白并抑制 HIV 复制。
- 显性负性突变体的利用。某些反转录病毒重组分子编码一个突变 Rev 蛋白，RevM10。RevM10 和 RRE 结合，但随后不能装配将 RNA 运输出核所需要的多蛋白复合物。
- 核酶的利用。有几个研究组制造了编码核酶的反转录病毒。RRz2 是一个直接对抗 *tat* 调控区域锤头结构的核酶；另一种类型，发夹核酶，也被用来切割 HIV 基因组。
- 细胞内抗体的利用。编码细胞 scFv 内抗体的反转录病毒（节 21.3.4）用来试图失活 Tat 或 Rev 调节蛋白，或 gp160 外壳糖蛋白。

在 HIV 感染的早期，由于免疫系统努力破坏感染细胞，因此有大量淋巴细胞的更新。如果免疫应答足够强，病毒可能被遏制在这一阶段。除了努力开发疫苗外，也致力于修饰 T 细胞的工作，以便它们能更有效地杀死感染细胞。已设计反转录病毒使得 $CD8^+$ 的 T 淋巴细胞表达嵌合的 T 细胞受体，对 HIV 感染细胞以靶向其细胞毒性反应。

所有这些方法的详细资料在 NIH 数据库中可找到‘感染性疾病’（对于任何感兴趣的方案点击科研摘要）。在体外操作过的细胞通常表现为对 HIV 感染有很高的抵抗力，并已在几个实验中证实了在体内长效（几个月到 1～2 年）的骨髓移植物。未解决的问题是移植物是否能够以足够高水平出现并提供临床上有用的 HIV 抗淋巴细胞池。可能通过使用细胞毒性化学物质或放射线率先破坏患者已有的骨髓而获得高水平的移植物——但是如果对 AIDS 患者这样做将是致命的方法。

（王莉莉　译）

进一步阅读

NIH database of gene therapy trials: www4.od.nih.gov/oba/rac/clinicaltrial.htm

Templeton NS, Lasic DD (eds) (2000) *Gene Therapy: Therapeutic Mechanisms and Strategies.* Marcel Dekker, New York.

Treacy EP, Valle D, Scriver CR (2001) Treatment of genetic Disease. In: *Metabolic and Molecular Basis of Inherited Disease*, 8th Edn (eds CR Scriver, AL Beaudet, WS Sly, MD Valle). McGraw-Hill, New York

Wiley database of approved gene therapy protocols: www.wiley.co.uk/genetherapy/clinical

参考文献

Acsadi G, Dickson G, Love DR *et al.* (1991) Human dystrophin expression in mdx mice after intramuscular injection of DNA constructs. *Nature* **352**, 815–818.

Cavazzana-Calvo M, Hacein-Bey S, de Saint Basile G *et al.* (2000) Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**, 669–672.

Chamberlain JS (2002) Gene therapy of muscular dystrophy. *Hum. Mol. Genet.* **11**, 2355–2362.

Check E (2002) A tragic setback (News Feature). *Nature* **420**, 116–118.

Check E (2003) Second cancer case halts gene-therapy trials (News Feature). *Nature* **421**, 305.

Costa T, Scriver CR, Childs B (1983) The effect of mendelian disease on human health: a measurement. *Am. J. Med.*

Genet. **21**, 231–242.

Daley GQ (2002) Prospects for stem cell therapeutics: myths and medicines. *Curr. Opin. Genet. Dev.* **12**, 607–613.

Davies JC, Geddes DM, Alton EWFW (2001) Gene therapy for cystic fibrosis. *J. Gene Med.* **3**, 409–417.

Dean W, Santos F, Stojkovic M *et al.* (2001) Conservation of methylation reprogramming in mammalian development: Aberrant reprogramming in cloned embryos. *Proc. Natl Acad. Sci. USA* **98**, 13734–13738.

Doudna JA, Cech TR (2002) The chemical repertoire of natural ribozymes. *Nature* **418**, 222–228.

Ferrari G, Gusella-De Angelis G, Coletta M *et al.* (1998) Muscle regeneration by bone marrow-derived myogenic progenitors. *Science* **279**, 1528–1530.

Fink DJ, Glorioso JC (1997) Engineering herpes simplex virus vectors for gene transfer to neurons. *Nature Med.* **3**, 357–359.

Galderisi U, Cascino A, Giordano A (1999) Antisense oligonucleotides as therapeutic agents. *J. Cell. Physiol.* **181**, 251–257.

Gardner MJ, Shallom SJ, Carlton JM *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511.

Goncz KK, Colosimo A, Dallapiccola B *et al.* (2001) Expression of F508 CFTR in normal mouse lung after site-specific modification of CFTR sequences by SFH. *Gene Ther.* **8**, 961–965.

Gordon JW (1999) Genetic enhancement in humans. *Science* **283**, 2023–2024.

Hacein-Bey-Abina S, Le Deist F, Carlier F *et al.* (2002) Sustained correction of X-linked severe combined immunodeficiency by *ex vivo* gene therapy. *New Engl. J. Med.* **346**, 1185–1193.

Hoogenboom HR, De Bruine AP, Hufton SE, Hoet RM, Arends J-W, Roovers RC (1998) Antibody phage display technology and its applications. *Immunotechnology* **4**, 1–20.

Hoppe-Seyler F, Butz K (2000) Peptide aptamers: powerful new tools for molecular medicine. *J. Mol. Med.* **78**, 426–430.

Huang L, Li S (1997) Liposomal gene delivery: a complex package. *Nature Biotechnol.* **15**, 620–621.

Hudson PJ (1999) Recombinant antibody constructs in cancer therapy. *Curr. Opin. Immunol.* **11**, 548–557.

Kafri T, Morgan D, Krahl T *et al.* (1998) Cellular immune response to adenoviral vector infected cells does not require *de novo* viral gene expression: implications for gene therapy. *Proc. Natl Acad. Sci. USA* **95**, 11377–11382.

Kay MA, Glorioso JC, Naldini L (2001) Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. *Nature Med.* **7**, 33–40.

Lindstedt S, Holme E, Locke EA *et al.* (1992) Treatment of hereditary tyrosinaemia type 1 by inhibition of 4-hydroxyphenylpyruvate dioxygenase. *Lancet* **340**, 813–817.

Marasco WA (1997) Intrabodies: turning the humoral immune system inside out for intracellular immunization. *Gene Ther.* **4**, 11–15.

Miyoshi H, Blomer U, Takahashi M, Gage FH, Verma IM (1998) Development of a self-inactivating lentivirus vector. *J. Virol.* **72**, 8150–8157.

Reyes-Sandoval A, Ertl HC (2001) DNA vaccines. *Curr. Mol. Med.* **1**, 217–243.

Russell CS, Clarke LA (1999) Recombinant proteins for genetic disease. *Clin. Genet.* **55**, 389–394.

Somia N, Verma IM (2000) Gene therapy: trials and tribulations. *Nat. Rev. Genet.* **1**, 91–99.

Stöger E, Vaquero C, Torres E *et al.* (2000) Cereal crops as viable production and storage systems for pharmaceutical scFv antibodies. *Plant Mol. Biol.* **42**, 583–590.

Sullenger BA, Gilboa E (2002) Emerging clinical applications of RNA. *Nature* **418**, 252–258.

Thomson JA, Itskovitz-Eldor J, Shapiro SS *et al.* (1998) Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145–1147.

Tse E, Rabbitts TH (2000) Intracellular antibody-caspase-mediated cell killing: an approach for application in cancer therapy. *Proc. Natl Acad. Sci. USA* **97**, 12266–12271.

Velander WH, Lubon H, Drohan WN (1997) Transgenic livestock as drug factories. *Sci. American* **276**, 70–74.

Vigna E, Naldini L (2000) Lentiviral vectors: excellent tools for experimental gene transfer and promising candidates for gene therapy. *J. Gene Med.* **5**, 308–316.

Weissman IL (2002) Stem cells – scientific, medical and political issues. *New Engl. J. Med.* **346**, 1576–1579.

White RR, Sullenger BA, Rusconi CP (2000) Developing aptamers into therapeutics. *J. Clin. Invest.* **106**, 929–934.

Wolf CR, Smith G, Smith RL (2000) Pharmacogenomics. *Br. Med. J.* **320**, 987–990.

Ye X, Al-Babili S, Kloti A *et al.* (2000) Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* **287**, 303–305.

Yeh P, Perricaudet M (1997) Advances in adenoviral vectors: from genetic engineering to their biology. *FASEB J.* **11**, 615–623.

词　汇　表

除此主词汇表外，还有四个专用词汇表：

PCR 方法词汇表，框 5.1，基因组学词汇表，框 8.1，

核酸杂交词汇表，框 6.3，后生动物种系发生群词汇表，框 12.5，

受累同胞对分析 [affected sib pair (ASP) analysis]：一种基于检测患有相同疾病的同胞间共享单体型的非参数连锁分析方法。见图 15.3。

等位基因 (allele)：同一基因的不同形式。

等位基因特异的寡核苷酸 (allele-specific oligonucleotide，ASO)：一段合成的通常为 20nt 长的寡核苷酸，在适当条件下与其靶序列的杂交可被单个碱基对的错配所破坏。ASO 可用作等位基因特异性杂交探针（图 6.11）或等位基因特异性 PCR 引物。见 *ARMS*。

等位基因关联性 (allelic association)：见连锁不平衡。

等位基因排斥 (allelic exclusion)：B 淋巴细胞两个免疫球蛋白等位基因或者 T 淋巴细胞两个 T 细胞受体等位基因中只有一个等位基因被表达的机制。更广义地是指任何自然发生的引起仅一个等位基因被表达的机制。见框 10.4。

等位基因异质性 (allelic heterogeneity)：在一个基因座上存在许多不同的致病等位基因。是某一基因功能丧失引起疾病的常见情况。见例图 16.1。

选择性剪接 (alternative splicing)：不同套剪接点序列的自然利用，可从单一基因产生一种以上的产物。见节 10.3.2，框 10.4。

Alu 重复 (Alu repeat)：在灵长类基因组中发现的一种高度重复的非编码 DNA 序列。

Alu-PCR：见框 5.1。

羊膜 (amnion)：哺乳动物 4 层胚外膜之一。见框 3.8。

非整倍性 (aneuploidy)：从一套完整的整倍体中增加或丢失一条或多条染色体的染色体构成。

复性 (anneal)：见框 6.3。

遗传早现 (anticipation)：一种疾病的严重程度在连续世代中增加的趋势。通常是由于调查偏倚（节 4.3.3）所致，但真正见于动态突变（节 16.6.4）。

反密码子 (anticodon)：tRNA 分子中与 mRNA 密码子碱基配对的 3 碱基序列。见图 1.7B 和图 1.20。

反效等位基因 (antimorph)：一个具有显性负效作用的等位基因。

反义 RNA (antisense RNA)：一个与正常 mRNA 互补的转录物，利用一个基因的非模板链生成。自然发生的反义 RNA 是基因表达的重要调节子。

反义链（模板链）[antisense strand (template strand)]：在转录过程中被 RNA 聚合酶用作模板来合成 mRNA 的某个基因的 DNA 链。见图 1.12。

细胞凋亡（apoptosis）：程序化细胞死亡。

藻类（archaea）：表面上类似细菌，但具有生命第三界分子特征的单细胞原核生物。见框 12.4。

扩增受阻突变系统（amplification refractory mutation system，ARMS）：等位基因特异性 PCR。见图 5.4 和图 18.10。

关联（association）：两个性状（疾病，标记等位基因等）以非随机频率共同发生的趋势。关联是一个简单的统计学观察，而不是一种遗传现象，但有时可由连锁不平衡引起。见节 15.4.1。

选型婚配（assortative mating）：具有相似表型或基因型的人之间的婚配（如高个者倾向于和高个者结婚，聋哑人倾向于和聋哑人结婚；有些人更愿与亲属结婚）。选型婚配在一个群体中能够引起基因型的非（哈迪-温伯格 Hardy-Weinberg）分布。

常染色体（autosome）：除性染色体 X、Y 之外的任何染色体。

同合性（autozygosity）：在近亲繁殖个体中，传递一致性的等位基因的纯合性。

同合性定位（autozygosity mapping）：对于常染色体隐性疾病，在大的近婚家族中寻找所有受累个体的相同等位基因为同合性的基因座。

细菌人工染色体（bacterial artificial chromosome，BAC）：能够插入长度不超过 300kb 的片段、在细菌中增殖的重组质粒。见节 5.4.3。

碱基互补性（base complementarity）：见框 6.3。

Bayesian 统计学（Bayesian statistics）：为多数遗传风险评估提供依据的统计学分枝。见框 18.4，图 18.15。

调查偏倚（bias of ascertainment）：由病例收集方式所造成的数据集中表型的歪曲比例。见节 15.2.1。

生物测量学（biometrics）：数量性状的统计学研究。

重亚硫酸盐测序（bisulfite sequencing）：检测 DNA 甲基化模式的一种方法。见节 18.3.6。

二价体（bivalent）：在减数分裂前期 I 可见的四链结构，由两条联会的同源染色体组成。见图 2.11。

BLAST：在序列数据库中检索与查询序列相匹配者的一组程序。见框 7.3。

胚泡（blastocyst）：胚胎发育的一个非常早期阶段，此时的胚胎由一个中空细胞球组成，含有一个内部充满液体的空腔（即胚泡腔）。

平端（blunt-ended）：DNA 片段的末端，无单链延伸。

自举法（bootstrapping）：为检查由比较序列分析构建的进化树的准确性而设计的一种统计方法。这种方法涉及用来自原始数据的随机次级样本替代一些原始序列数据并重新计算每一循环中原始树正确似然性的大量循环。见图 12.18。

边界元件（boundary element）：染色体中用来确定共调控染色质区域边界的序列。见框 8.2。

分支点（branch site）：在 mRNA 加工过程中，位于剪接受体上游 10～50 个碱基，包含形成套索剪接中间体所需腺苷的一段不甚明确的序列（一致序列为 CTRAY：R＝嘌呤，Y＝嘧啶）。见图 1.15。

C 值颠倒现象（C value paradox）：一个有机体细胞内的 DNA 含量（C 值）与该有机体的复杂性缺乏直接关系。

候选基因（candidate gene）：在定位克隆中，位于疑似存在疾病基因的适当染色体位置上的基因。通过在患者中筛查突变来验证。

帽（cap）：细胞添加的、封闭 mRNA 5′端的一种特殊的化学基团。见图 1.17。

互补 DNA（complementary DNA，cDNA）：通过实验（图 4.8、图 6.5）手段或者在体内（图 7.13 和图 14.18）以 mRNA 为模板由反转录酶合成的 DNA。

cDNA 选择（cDNA selection）：一种以杂交为基础的获取 cDNA 文库中相应的基因组克隆的方法。见图 7.11。

厘摩（centiMorgan，cM）：遗传距离单位。在减数分裂过程中，相距 1cM 的基因座间有 1%重组的可能性。遗传距离和物理距离间的关系见图 13.4。

厘伦琴（centiRay，cR）：见框 8.1。

着丝粒融合（centric fusion）：见罗伯逊融合（Robertsonian fusion）。

着丝粒（centromere）：染色体的主缢痕，分隔长臂和短臂，也是细胞分裂过程中，纺锤丝附着以牵引染色单体分离的位点。见节 2.3.2。

人类多态研究中心家系（CEPE family）：由巴黎 Centre d'Etude du Polymorphisme Humain 收集的，用以协助制作标记一标记框架图的一套家系。

错配化学裂解（chemical cleavage of mismatch，CCM）：筛查 kb 大小 DNA 片段突变的一种方法。见表 18.2，图 18.4B。

交叉（chiasma，复数：chiasmata）：显微镜下可见的减数分裂重组的现象。见图 13.3。

嵌合体（chimera）：来源于不止一个合子的有机体。见图 4.10。

脊索动物（chordate）：见框 12.5。

绒毛膜（chorion）：哺乳动物四层胚外膜之一。见框 3.8。

染色单体（chromatid）：从细胞周期的 S 期末直到细胞分裂后期，染色体由两条姐妹染色单体组成。各自含有一条完整的双螺旋，二者是彼此的精确拷贝。

染色质纤维（chromatin fiber）：由 DNA 和组蛋白组成的 30nm 的螺线管，被认为是染色质的基本构象。

染色体涂染（chromosome painting）：用 FISH 方法进行整条染色体的荧光标记，以来自单一染色体的许多不同 DNA 序列的混合物为探针。

染色体步查（chromosome walking）：通过筛查基因组文库中与特征性克隆部分重叠的克隆来找出染色体上与其邻近的序列。

顺式作用（*cis*-acting）：作为调节因子，只有当其是同一 DNA 分子或染色体的一部分时，才能调控一个基因的活性。与反式作用调节因子相对，后者可调控其靶序列而与染色体位置无关。

克隆（clone）：见框 8.1。

溯祖时间（coalescence time）：群体遗传学中，回溯至最近共同祖先的世代数。见框 12.6。

编码 DNA（coding DNA）：编码某一多肽的氨基酸序列（或偶尔为不翻译成多肽的功能性成熟 RNA）的 DNA。

密码子（codon）：指定一个氨基酸或翻译终止信号的核苷酸三联体（严格意义上指 mRNA 序列，但可引申至基因组编码 DNA）。

选择系数（coefficient of selection）：相比最成功的基因型，某一基因型不能繁衍的几率。见节 4.5.2。

比较基因组杂交（comparative genomic hybridization，CGH）：运用竞争性荧光原位杂交检测来扩增或缺失的染色体区域，特别是在肿瘤中。见图 17.3。

互补链（complementary strand）：如果两条核酸链能形成足够多的碱基对以产生稳定的双链结构，就被称为在序列上是互补的。

互补作用（complementation）：如果两个等位基因在一起时，可恢复野生型表型，则二者是互补的（框 4.2）。等位基因互补通常仅发生在其位于不同的基因座时，但一些等位基因之间的互补作用亦可发生。

复杂性（complexity）：一个基因组的复杂性是指单一序列的总长度或比例。低复杂性序列在一个基因组或样本中会多次出现。

复合杂合子（compound heterozygote）：一个人在某一基因座上具有两个不同的突变等位基因。

条件敲除（conditional knock-out）：在某些条件（如温度升高）下或某些而不是其他细胞中人为设计的能引起基因功能丧失的突变体。

同线性保守（conservation of synteny）：两种有机体遗传图比较时，如果一种生物中位于同一条染色体上的两个或更多的基因座在另一种生物中也位于单一染色体上，则为同线性保守。

保守型置换（conservative substitution）：造成一个密码子被另一个编码不同氨基酸的密码子所替代的突变，后者在化学性质上与原来的氨基酸有关。

体质性（constitutional）：一种曾存在于受精卵中的基因型、异常或突变，因而存在于个体的所有细胞中，区别于体细胞改变。

组成型表达（constitutive expression）：指基因持续激活的一种状态。引起不适当组成型表达的突变通常具有致病性。

叠连群（contig）：一个表或图，表示共同含有一条原始的连续 DNA 链序列的克隆重叠片段的有序排列。

邻接基因（节段性非整倍体）综合征 [contiguous gene (segmental aneuploidy) syndrome]：由邻接的一组基因的缺失引起的综合征，其中几个或全部基因与表型有关。

连续性状（continuous character）：每个人都具有的但程度不同的性状，如身高——与双歧性状如多趾症（一些人有而另一些人没有）相区别。

CpG 二核苷酸（CpG dinucleotide）：一个较长 DNA 分子中的 5′CG 3′序列。CpG 二核苷酸是哺乳动物所特有的 DNA 甲基化系统的靶标，对基因表达调控起重要作用。

CpG 岛（CpG island）：一段短的 DNA 序列，通常小于 1kb、包含常见（密集）的非甲基化 CpG 二核苷酸。CpG 岛往往是基因 5′端的标记。见框 9.3。

Cre-lox 系统（Cre-lox system）：一种能够产生预先确定的染色体缺失的一种技术。*Cre* 是一个其产物能促进 *lox*P 序列之间重组的噬菌体 P1 基因。如此称呼是因为它能产生重组，见图 20.10 和图 20.11。

隐蔽剪接位点（cryptic splice site）：前体 mRNA 中存在的与剪接位点具有某些同源性的序列。当剪接受到干扰时或碱基置换突变增加了与正常剪接位点的相似性后，隐蔽剪接位点可作为剪接位点使用。见图 11.12 和图 11.13。

细胞骨架（cytoskeleton）：存在于细胞内，在调节细胞形态、细胞运动和细胞内转运方面具有极其重要作用的蛋白质网状骨架。见框 3.2。

简并寡核苷酸（degenerate oligonucleotide）：在某些核苷酸位置上由于灵活性而平行合成的一组寡核苷酸，用作杂交探针或 PCR 引物时能够杂交或扩增一系列序列。

变性（denaturation）：互补双链解离为单链 DNA 和/或 RNA。

后口动物（deuterostome）：见框 12.5。

变性梯度凝胶电泳（denaturing gradient gel electrophoresis，DGGE）：将 PCR 产物通过具有变性梯度的凝胶电泳来检测突变的方法。见图 18.3B。

变性高效液相色谱（denaturing high-performance liquid chromatography，dHPLC）：利用杂合双链的性质来检测突变的一种方法，具有高灵敏度和高通量的特点。见图 18.4。

双歧性状（dichotomous character）：一些人有但另一些人没有的性状，如多趾症——与连续性状如身高（每个人都有但程度不同）相对。

差别显示（differential display）：见框 5.1。

分化（differentiation）：细胞开始特化并决定最终形成成熟的特殊细胞类型的过程。

二倍体（diploid）：具有两个拷贝的每种类型染色体，是大多数人类体细胞的正常组成。

（染色体的）远侧 [distal（of chromosomes）]：距染色体着丝粒相对较远的位置。

DNA 芯片（DNA chip）：见微阵列。

DNA 指纹（DNA fingerprinting）：一种已过时的以法律或法医为目的，基于运用高度可变的小卫星探针进行 Southern 印迹分析来鉴定个体的方法。见图 18.19。

DNA 文库（DNA library）：见框 8.1。

DNA 标记（DNA marker）：见标记（遗传）。

DNA 甲基化（DNA methylation）：通常指人类 DNA 中胞嘧啶（常位于 CpG 二核苷酸处）转变为 5-甲基胞嘧啶。

DNA 谱（DNA profiling）：利用一系列多态位点不同的基因型来识别个体，通常是为了伦理或法医的目的。见节 18.7。

DNA 酶 Ⅰ 高敏感位点（DNAse Ⅰ-hypersensitive site）：可被 DNA 酶 Ⅰ 迅速消化的染色质区。被认为是重要的远程控制序列。见节 10.5.2。

显性负效（反效等位基因）[dominant negative（antimorph）]：在杂合子中基因产物能够抑制野生型基因产物功能的突变基因，例如，见图 16.4。

显性（dominant）：人类遗传中，在杂合子中能够表达的任何性状。也见半显性。

剂量补偿（dosage compensation）：不同数量的基因能产生等量产物的任何体系。在哺乳动物，它描述确保 XX 和 XY 细胞具有等量的 X 染色体编码基因产物的 X—失活机制。见图 10.26。

剂量敏感性（dosage sensitivity）：一个基因拷贝数量的改变引起表型异常的特性。

复制（拷贝或重复的）转座 [duplicative（or copy or replicative）transposition]：一个

DNA 序列的拷贝发生转座，而其原序列位置不变。

动态突变（dynamic mutation）：一种不稳定的，在亲代和子代间改变长度的扩增重复序列，见节 16.6.4。

外胚层（ectoderm）：胚胎三个胚层之一。在原肠胚形成期由上胚层细胞形成（图 3.15），将发育为神经系统和外层上皮（框 3.5）。

电穿孔（electroporation）：在体外利用瞬时高压脉冲将 DNA 转入细胞内的方法。

胚胎干细胞（embryonic stem cell）：见 ES 细胞。

经验风险率（empiric risk）：由调查数据而不是遗传理论计算的风险率。大多数非孟德尔疾病的遗传咨询是以经验风险率为基础的。见节 4.4.4。

内胚层（endoderm）：胚胎三个胚层之一。在原肠胚形成期由上胚层迁移出来的细胞形成（图 3.15）。胚胎内胚层的衍生物见框 3.5。

增强子捕获（enhancer trap）：鉴定某一生物体内强表达基因的技术。

增强子（enhancer）：一组能够促进基因转录的短序列元件，其功能并不严格地依赖于其精确的定位和方向。见框 10.2。

上胚层（epiblast）：在前原肠胚形成期胚胎中，将发育为正确胚胎全部三个胚层及胚外外胚层和胚外中胚层的细胞层。与下胚层对应。

表观遗传的（epigenetic）：可遗传的（从母细胞到子细胞，或有时从亲代到子代）但不是 DNA 序列改变引起的。DNA 甲基化是最适合理解表观遗传的机制。

附加体（episome）：细胞中能够以自主的染色体额外形式存在的任意 DNA 序列。通常用来描述 DNA 的自我复制和染色体额外形式。

表位（epitope）：抗原上与特定抗体发生反应的部分。

胚胎干细胞（ES cell，embryonic stem cell）：来源于一个胚胎的未分化、具有多能性的细胞。是遗传操作的重要工具。见节 3.4.5，图 21.1 和图 20.3。

表达序列标签（expressed sequence tag，EST）：见框 8.1。

常染色质（euchromatin）：包含有转录活性 DNA 的核基因组部分。与异染色质不同，具有相对伸展的构象。

优生学（eugenics）：通过从最佳类型中进行选择性生育（正优生学）或阻止不合需要的类型生育（负优生学）以“改善”群体质量。

整倍性（euploidy）：无额外或丢失的、具有一套或多套完整染色体的状态；非整倍性的对立面。

排除定位（exclusion mapping）：具有阴性结果的遗传定位。能显示要定位的基因座并不定位于某一特殊位置。尤其有利于无需实验室的突变筛查就可排除一个可能的候选基因。

外显子捕获（exon trapping）：在一专门的载体内，从一个克隆的基因组 DNA 中鉴别检定能够剪接成外显子的序列的技术。见图 7.10。

外显子（exon）：指在成熟 RNA 产物中基因的一个片段。单个外显子可含有编码 DNA 和/或非编码 DNA（非翻译序列）。见图 1.14 和图 1.19。

表达文库（expression library）：一个克隆至载体并可以表达的 cDNA 文库。见节 5.6。

表达谱（expression profiling）：获得 mRNA 水平全部基因组图谱，通常由全部细胞

cDNA 的微阵列分析所得（图 17.18，19.8）。也见 *SAGE*。

细胞外基质（extracellular matrix）：细胞外空间可见的与细胞表面基底膜相连的网状物质。它提供细胞黏附的支架并用于促进细胞的增殖。

适合度（fitness，f）：群体遗传学中，将基因型成功传递给下一代的衡量标准。也称为生物适合度或生殖适合度。f 值总是介于 0～1 之间。

荧光原位杂交（fluorescence *in situ* hybridization，FISH）：应用荧光标记 DNA 或 RNA 探针进行的原位杂交。现代分子遗传学的重要技术——见图 2.17 和图 2.18。

折叠（fold）：三维蛋白结构模型。大多数蛋白质结构是由有限的，在许多蛋白质中共享的折叠组成。见节 19.4.4。

建立者效应（founder effect）：群体中某一特定等位基因的高频率，因为该群体是从少数建立者衍生而来，其中一个或多个人携带该等位基因。

移码突变（frameshift mutation）：由于增加或缺失非 3 的倍数的若干碱基改变了mRNA 正常的翻译可读框的突变。

功能基因组学（functional genomics）：通过对大量基因、甚至基因组所有基因的表达/功能的平行聚类分析而进行的大规模的基因功能分析。

融合基因（fusion gene）：含有来源于两个不同基因的编码序列的基因，通常由不等交换（图 9.17）或染色体易位（图 18.7A）引起。

融合蛋白（fusion protein）：天然的或人工的融合基因的产物。一条单一的多肽链含有正常的两条或更多条独立多肽部分的氨基酸序列。见框 20.2。

原肠胚形成（gastrulation）：涉及由两层的前原肠胚胚胎（由上胚层和下胚层组成）转变成含有三胚层（外胚层、中胚层和内胚层）胚胎的高度动态过程。

基因转变（gene conversion）：自然发生的不可逆的遗传改变，即一条 DNA 链的序列改变后与另一条 DNA 链上的序列一致。见图 11.10。

基因频率（gene frequency）：一个基因座上所有涉及到的等位基因的比例，见节 4.5。实际上我们指的是等位基因频率，但现在基因频率的使用太常见而无法改变。

基因替代（gene replacement）：用一个有正确功能的基因替代内源性的异常基因的基因治疗。见图 21.4。

基因添加/增加（gene supplementation/augmentation）：不需对任何内源性基因进行操作而将功能基因引入患者细胞的基因治疗。可适当纠正功能丧失表型。见图 21.4。

基因打靶（gene targeting）：细胞或生物体内基因的靶向修饰。见图 20.7，图 20.8，图 20.10。

基因跟踪（gene tracking）：使用连锁的标记追踪染色体片段来预测家系内的基因型。有时称间接检验。见框 18.3。

基因捕获（gene trap）：鉴别选择已插入基因内的转基因插入片段的方法。

遗传距离（genetic distance）：遗传图上的距离，以重组值和作图函数来定义，以厘摩为衡量标准。见节 13.1。

遗传增强（genetic enhancement）：通过某些有益的方法改变一个正常人表型的分子遗传学技术的可能应用。见 21 章伦理学框 3。

遗传图（genetic map）：见框 8.1。

遗传丰余 (genetic redundancy)：多个基因座的基因平行发挥相同的功能，以至某一基因座功能丧失突变不会引起完全性功能丧失。

基因组浏览器 (genome browser)：为查询基因组数据库提供图形接口的程序。见节 8.3.6。

基因组 (genome)：一个细胞器、细胞或生物体内不同 DNA 的总和。人类基因组包括 25 条不同的 DNA 分子，即线粒体 DNA 分子加上 24 条不同的染色体的 DNA 分子。参照转录物组、蛋白质组。

全基因组 p 值 (genome-wide *p* value)：检验连锁或关联时，在全基因组扫描任何位置观察到被涉及的统计量这一无效假设的概率（参照逐点 p 值）。见节 15.3.4。

基因型 (genotype)：指一个体的整体或某一特定基因座的遗传组成。

种质（生殖）细胞（或配子） (germ cell or gamete)：精细胞和卵细胞。

生殖嵌合体 (germinal mosaic)：一个个体具有携带其他生殖系细胞中未见的突变的生殖系细胞亚群。见图 4.9。

种（生殖）系 (germ-line)：种质（生殖）细胞和形成种质（生殖）细胞的细胞及构成机体的其他细胞。

单倍体 (haploid)：描述仅含有每条染色体单一拷贝的细胞（典型的是配子，例如人类精子和卵子中的 23 条染色体）。

单倍剂量不足 (haploinsufficiency)：如果产生某一正常表型需要比单一拷贝产生的量更多的基因产物，则此基因座表现出单倍剂量不足。见节 16.4.2，表 16.2。

单体型 (haplotype)：在一条染色体上紧密相连的基因座的系列等位基因。

单体型板块 (haplotype block)：在一个群体许多成员中存在的多态的延伸的染色体 DNA 片段（代表性的是 10～100kb）的特殊变异体，可能是一个祖先变异体。见框 12.6 和节 15.4.3。

哈迪-温伯格分布 (Hardy-Weinberg distribution)：在某一条件下，一个群体中基因频率和基因型频率之间的一种简单关系。见节 4.5。

半合子的 (hemizygous)：在二倍体细胞中仅有一个基因或 DNA 序列的单一拷贝。对于大多数性染色体基因来说，雄性是半合子的。某一常染色体上的缺失可导致雄性和雌性的半合子性。

遗传率 (heritability)：遗传因素引起的某一性状形成的比率。见框 4.4。

异染色质 (heterochromatin)：在整个细胞周期中呈高度浓缩状态，没有或未发现活性基因表达的染色质区域。组成型异染色质多见于着丝粒及其他一些区域。见图 2.15。

异源双链体 (heteroduplex)：两条链间有一些错配的双链 DNA，在突变检测中很重要。见节 18.3.2。

异质性 (heteroplasmy)：常位于单一细胞内的镶嵌现象，如线粒体 DNA 变异体。见节 4.2.5 和节 11.4.2。

杂合子优势 (heterozygote advantage)：指有某种突变的杂合子个体较野生型纯合子个体具有生殖优势的现象，有时也称超显性。杂合子优势是几种严重的隐性疾病仍旧常见的原因。见框 4.8。

杂合子 (heterozygote)：指在某一特定基因座上具有两个不同等位基因的个体。

同源染色体（homologs chromosome）：指一个二倍体细胞内染色体的两个拷贝。与姐妹染色单体不同，同源染色体不是彼此的拷贝：一条遗传自父方，另一条遗传自母方。

同源基因（homologs gene）：由于物种间（种间同源）或物种内（种内同源）密切的进化关系造成序列显著性相关的两个或多个基因。

同质性（homoplasmy）：指一个细胞或生物体含有相同线粒体 DNA 的所有拷贝。参照异质性。

纯合子（homozygote）：在某一特定基因座上具有两个相同等位基因的个体。临床上，在某一基因座上有两个正常功能等位基因的人常被认为是纯合性 AA，或者有两个病理性等位基因的人被认为是纯合性 aa，而不考虑等位基因在 DNA 序列水平实际上是否完全相同。具有传递一致性的等位基因的纯合性称同合性。

热点（hotspot）：与重组或突变的异常高频率相关的序列。

同源框基因（hox gene）：指成簇存在的、在前-后轴成型过程中起重要作用的同源框基因的一组亚群。见图 3.10 和图 3.12。

杂种细胞嵌板（hybrid cell panel）：用于物理作图的体细胞杂种或辐射杂种细胞的集合。

超效等位基因（hypermorph）：一个引起产物数量或活性增加的等位基因。

下胚层（hypoblast）：在前原肠胚形成期胚胎中将发育为胚外内胚层的细胞层。

亚效等位基因（hypomorph）：一个能引起产物数量或活性降低的等位基因。

传递一致性（identity by descent，IBD）：一个或两个个体中已知是遗传自一个可证实为共同祖先一致的等位基因。

状态一致性（identity by state，IBS）：看起来是一致的，但由于没有可证实的共同来源而可能具有或可能不具有传递一致性的等位基因。见图 15.2。

原位杂交（*in situ* hybridization）：分子杂交的一种方法，靶核苷酸可以是制备好的染色体中变性的 DNA（**染色体原位杂交**），或者是固定于显微镜玻片的组织切片中的细胞 RNA（**组织原位杂交**，图 6.15），或者是整个胚胎内的 RNA（**整装原位杂交**，图 7.15）。

印记（imprinting）：由其亲代起源决定一个基因的表达。见节 10.5.4 和框 16.6。

近交（inbreeding）：与有血缘关系的亲属婚配。此定义是相对的，因为根本上每个人都有亲缘关系。近交系数是指一个人具有传递一致性基因的比率。

诱导型启动子（inducible promoter）：其活性可被某种外部因素激活或抑制的启动子。见图 20.5A。

提供信息的减数分裂（informative meiosis）：连锁分析中，如果从系谱基因型可以推断出是否为重组体（对于给定的一对基因座），我们说此减数分裂是可提供信息的。见框 13.2。

内细胞团（inner cell mass）：位于囊胚内部的能够发育为正常胚胎的一群细胞。见图 3.13。

插入诱变（insertional mutagenesis）：由于基因内一无关 DNA 序列的插入所引起的基因突变（通常导致基因功能的丧失）。

干涉（interference）：在减数分裂中，某一交换可抑制相同染色体区域内进一步发生交换的趋势。见节 13.1.3。

间期（interphase）：细胞周期中细胞未发生分裂的所有时期。

内含子（intron）：指基因内分隔相邻外显子的非编码 DNA 序列。在基因表达过程中，内含子转录至 RNA，但随后通过剪接，内含子序列从前体 mRNA 中被除去。见图 1.14。可依据剪接机制（框 12.1）或者当其分隔编码 DNA 序列时通过它们在密码子上的精确位置（框 12.2）进行分类。

等臂染色体（isochromosome）：由两条相同长度的臂构成的异常的对称性染色体，通常含有一条正常染色体的长臂或短臂。

异构体/同工酶（isoform/isozyme）：一种蛋白质/酶的两种不同形式。

同系的（isogenic）：具有相同基因型的两个或更多的生物体或细胞。例如，属于某一特定近交系如 C57B10/J 的不同鼠是同系的。

核型（karyotype）：严格地讲，核型是指一个细胞或个体的染色体构成的总和，如 46，XY。但此术语通常也泛指一个细胞内按次序成对排列分类的染色体图，如图 2.14。

敲倒（knock-down）：运用某种方法，例如特异的反义 RNA 或 siRNA 结合 RNA 转录物来实现基因表达的靶向抑制。

敲入突变（knock-in mutation）：通过导入活性基因替代某一活性基因的靶向突变。见图 20.14。

敲除突变（knock-out mutation）：在一完整细胞内某一基因的靶向失活。

后随链（lagging strand）：DNA 复制过程中以冈崎片段方式合成的链。见图 1.9。

前导链（leading strand）：DNA 复制过程中连续合成的链。见图 1.9。

连接（ligation）：两个分子末端的核苷酸（分子间连接）或同一分子两个末端的核苷酸（分子内连接、环化）之间 3′-5′磷酸二酯键的形成。

长散在核元件（long interspersed nuclear element，LINE）：约占人类基因组 20%的一类重复 DNA 序列（表 9.15）。其中一些为活性转座元件。见图 9.17 和图 9.18。

谱系（lineage）：来源于一个始祖细胞的一组细胞。

连锁不平衡（linkage disequilibrium）：位于分隔的但紧密连锁的基因座上特定等位基因间的统计学关联。通常是由被研究群体中常见的特定祖先单体型的结果。是高分辨作图的重要工具。见图 13.10、图 15.6 和框 15.2。

连锁（linkage）：某些性状（表型、标记等位基因等）因其决定因素在某一特定染色体上紧密相邻而在系谱中共分离的趋势。

接头寡核苷酸（linker oligonucleotide）：能够与感兴趣的 DNA 分子连接的双链寡核苷酸，被设计含有某些需要的特征，如有利的限制酶位点。

脂质体（liposome）：人工合成的、用于将感兴趣的分子转入细胞的脂质小囊泡。见图 21.8。

基因座控制区（locus control region，LCR）：含有可调控一个基因簇内可能相距几十 Kb 的基因表达调控元件的 DNA 序列。见图 10.23。

基因座异质性（locus heterogeneity）：同一疾病或表型可由不同基因座的突变所引起。是遗传病定位中一个常见问题。见节 4.2.4 和节 16.7.2。

基因座（locus）：决定一个基因或一段 DNA 序列位置的独一无二的染色体定位。

对数优势比〔lod score（z）〕：基因座间遗传连锁可能性的衡量标准。当重组值为 θ 时，基因座是连锁的而不是非连锁的优势对数（以 10 为底）。对于孟德尔性状，对数优势比大于＋3，肯定连锁的存在；如小于－2，则否认连锁的存在。见框 13.3 和图 13.7。

杂合性丢失（loss of heterozygosity，LOH）：当本质上的基因型是杂合的时候，肿瘤细胞或其他体细胞的纯合性或半合性，体细胞遗传改变的证据。见图 17.6 和图 17.7。

基质辅助激光解吸离子化（matrix-assisted laser desorption/ionization，MALDI）：常用于鉴定 DNA 或蛋白分子的大规模质谱分析方法。见框 19.4。

显示杂合子（manifesting heterozygote）：一个 X 连锁隐性疾病的女性携带者可能由于不对称性 X-染色体失活而表现出某些临床症状的性状。见节 4.2.2。

多重扩增探针杂交（multiplex amplifiable probe hybridization，MAPH）：检测外显子缺失或复制的一种方法。见框 18.1。

作图函数（mapping function）：描述重组值和遗传距离关系的数学方程式。作图函数取决于干涉阻止相邻双重重组体发生的程度。见节 13.1.3。

标记染色体（marker chromosome）：额外的、来源不明的异常染色体。

遗传标记［marker（genetic）］：任何可用于在系谱中追踪某些染色体片段多态性的孟德尔特性。遗传标记通常指 DNA 多态性。见框 13.1。

母系遗传（matrilineal inheritance）：仅从母亲传递，但可传给任一性别的子女；线粒体的遗传方式。见图 4.4。

平均杂合度（mean heterozygosity）：对于一个随机选择的个体，某一标记为杂合的似然性。是标记对连锁分析作用的衡量标准。见节 11.2.2。

熔解温度（melting temperature，T_m）：见框 6.3。

孟德尔遗传（mendelian）：一种系谱模式，符合图 4.2.A 显示的原型之一。在分子遗传学家的观念中，一个性状只要由单一的染色体位置决定，则该性状呈现孟德尔遗传系谱模式，而无须考虑其决定因素是否为一个基因。

中胚层（mesoderm）：胚胎三个胚层之一，在原肠胚形成期由上胚层迁移出来的细胞形成。胚胎中胚层的衍生物见图 3.15 和框 3.5。

中期（metaphase）：细胞分裂（有丝分类或减数分裂）的一个阶段，此时染色体最大限度浓缩并排列于细胞赤道板（中期板）。见图 2.10、图 2.11 和图 2.15。

后生动物（metazoans）：与单细胞原生动物相对应的多细胞动物。

微阵列（microarray）：位于玻璃表面，用于杂交分析的不同 DNA 或寡核苷酸序列的小型阵列。制作 DNA 芯片所用的序列可以是自动化装置制备的 DNA 分子或原位合成的寡核苷酸序列。

微缺失（microdeletion）：非常小而无法在显微镜下观察到的染色体缺失（一般＜3Mb）。

微 RNA［miRNA（micro-RNA）］：在正常基因组内编码短的 RNA 分子（22nt），在基因表达调节中起作用。也可能是染色质结构的组成部分，有时称为小瞬时 RNA（stRNA）。见图 9.6 和图 20.12。

微卫星（microsatellite）：一段非常简单的 DNA 序列［通常 1～4bp，例如（CA）n］的短连续串联重复（一般小于 0.1kb）。通常是多态的，是 20 世纪 90 年代遗传定位的基本工具。有时也称为简单串联重复（STR）或简单短串联重复（SSR）多态性。见图 7.7 和图 7.8。

微卫星不稳定性（microsatellite instability）：某些肿瘤细胞特有的现象，DNA 复制过程中肿瘤细胞内微卫星的重复拷贝数易发生随机的变化。可缩写为 MIN，MSI 或 RER 复制错误。见图 17.7。

微管（microtubule）：由微管蛋白多聚体构成长的中空的圆管，是细胞骨架的组成部分。见框 3.2。

MIM 号（MIM number）：一个基因或孟德尔性状在 OMLM 数据库的目录号。

小卫星 DNA（minisatellite DNA）：一连串中等大小的短串联重复 DNA 序列，一般长 0.1～20kb。见框 7.2。高度可变的小卫星 DNA 是 DNA 指纹图和许多 VNTR 标记的基础。

小测序（minisequencing）：在一个预定的位置仅对一条引物下游 1 或 2 个碱基测序来检测序列变异体的一种方法。常用于微阵列的设立。见图 18.9B。

错配修复（mismatch repair）：自然发生的酶促过程，替换 DNA 双链中错误配对核苷酸（大多可能由于 DNA 复制过程中的错误所致），从而获得正确的 Watson-Crick 碱基配对。

错义突变（missense mutation）：导致氨基酸改变的核苷酸替换。见 Box11.3。

修饰基因（modifier gene）：其表达可影响另一个基因座突变产生的表型的基因。见节 16.6.3。

单等位基因表达（monoallelic expression）：在一个细胞内，一个基因的两个拷贝中仅有一个拷贝表达。原因有 X 失活，印记或其他表遗传的改变，以及免疫球蛋白基因和 T 细胞受体基因发生的基因重排。具体例子见框 10.4。

单克隆抗体（monoclonal antibody，mAb）：应用杂交瘤技术生产的具有专一性的纯化抗体。与免疫产生的多克隆抗体不同。见框 7.4。

嵌合体（mosaic）：来源于一单个合子，具有两个或多个遗传上不同的细胞系的个体。这种差异可以是点突变，染色体改变等。见图 4.10。

多因子的（multifactorial）：由一些不确定的遗传和环境因素所共同决定的某种性状。参照多基因的。

多基因家族（multigene family）：基因组内一组进化上相关的基因座，其中至少一个能编码功能性产物。见节 9.3。

新效基因（neomorph）：具有一种新活性或产物的等位基因。

神经嵴（neural crest）：一群具有高度多向性的细胞，能发育为周围神经系统、表皮黑色素组织、某些骨骼和肌肉，视网膜以及其他结构的一部分。神经嵴细胞作为一种特殊细胞群形成于神经胚形成期，沿着神经褶侧缘出现然后从神经板分离并迁移至机体内一些特定的位置。见图 3.16A。

不分离（nondisjunction）：细胞分裂后期染色体（有丝分裂或减数分裂Ⅱ期的姐妹染色单体或减数分裂Ⅰ配对的同源染色体）分离失败。是引起染色体数目异常的主要原

因。见节 2.5.2。

非同源重组（nonhomologous recombination）：在非同源序列或仅有部分同源性的序列间发生的重组。在遗传或染色体水平上，是引起插入或缺失的主要原因。见例图 11.7 和图 16.2。

非参数的（nonparametric）：连锁分析中一种不依赖于某种特定遗传模式的分析方法。如受累同胞对分析。

不外显（nonpenetrance）：指携带通常引起显性表型等位基因的个体但却不表现出相应表型的情况。多由于其他基因座影响或环境作用所致。是遗传咨询的一个陷阱，图 4.5B 所示的一个例子。

无义突变（nonsense mutation）：发生于密码子中使其改变为终止密码子的突变。见框 11.3。

无义突变介导的 mRNA 降解（nonsense-mediated mRNA decay）：降解包含提前出现的终止密码子（最末剪接点上游＞50nt）的 mRNA 分子的一种细胞机制。见节 11.4.4。

RNA 印迹（northern blot）：带有经凝胶电泳按片段大小分离的 RNA 分子的膜，用作杂交分析的靶标。用来检测成人或胎儿组织标本中感兴趣的某个基因转录物的大小。见图 5.13。

脊索（notochord）：在低等脊索动物、低等脊椎动物和比较复杂的脊椎动物胚胎中，形成机体支撑轴的一种柔软的棒状的结构。

核仁组织区（nucleolar organizer region，NOR）：人类 13，14，15，21 和 22 号染色体的卫星柄。NOR 含有大量核糖体 DNA 基因并能被选择性银染。每个 NOR 在细胞分裂末期形成一个核仁。核仁在间期融合。

核小体（nucleosome）：染色质的结构单位。见图 2.3。

无效等位基因（null allele）：不能产生产物的突变等位基因。

寡基因的（oligogenic）：由少量基因共同作用所决定的性状。

寡核苷酸连接测定法（oligonucleotide ligation assay，OLA）：一种检测预先确定的序列变化的方法。见图 18.11。

OMIM（On-line Mendelian Inheritance in Man）人类在线孟德尔遗传：人类基因和孟德尔性状最重要的数据库（http://www3.ncbi.nlm.nih.gov/omim/.）MIM 号是进入 OMIM 的索引号码。

癌基因（oncogene）：与控制细胞增殖有关的基因，当其过度激活时有助于使正常细胞转化为肿瘤细胞。见表 17.1。此词原意仅指这种基因的激活形式，而正常细胞中存在的基因称为原癌基因，但现在这种区别常被忽略。

一基因一酶假说（one gene-one enzyme hypothesis）：1941 年由 Beadle 和 Tatum 提出的假说：每一个基因的主要作用是决定一种酶的结构。此假说在历史上很重要，但现在被视为仅仅是基因功能范围的一部分。

可读框（open reading frame，ORF）：一条有意义的长 DNA 序列，其中至少一种可能的读框没有终止密码子。因为每条链能有三种读框，所以对于一 DNA 双链来说可能有 6 种读框。

种间同源（ortholog）：不同物种间一组同源基因中的一个（例如人类的 *PAX3* 基因；小鼠的 *pax3* 基因）。见框 12.3。

超显性的（overdominant）：显示杂合子优势的表型。用于群体遗传学的一个术语。

回文序列（palindrome）：一段 DNA 序列，诸如 ATCGAT，每条链按 5′→3′的方向阅读均相同的 DNA 序列。DNA-蛋白的识别，例如通过限制性内切酶常依赖回文序列。

臂内倒位（paracentric inversion）：不包含着丝粒的某一染色体片段的倒位。见图 2.20。

种内同源（paralog）：在单一物种内一组同源基因中的一个。见框 12.3。

参数的（parametric）：在连锁分析中，需要严格特定的遗传模式的方法，如标准的 LOD 值分析。

部分［酶切］消化（partial digestion）：通常指限制性内切酶对 DNA 的消化，消化在所有靶序列被切断之前停止。其目的是为了产生重叠片段。见图 5.9。

外显率（penetrance）：在一既定的表型，基因型表明自身的频率。

臂间倒位（pericentric inversion）：包含着丝粒的某一染色体片段的倒位。见图 2.20。

噬菌体展示（phage display）：将外源基因插入噬菌体载体并表达为在噬菌体表面展示的多肽的一种表达克隆方法。

药物遗传学（pharmacogenetics）：研究单个基因或等位基因对药物代谢或药物功能的影响。

药物基因组学（pharmacogenomics）：利用基因组资源（基因组序列，表达谱等）鉴定新的药物靶标。

连锁标记相态（phase of linked marker）：位于 2 个连锁基因座的等位基因间的关系（连接或排斥）。如果等位基因 A1 与等位基因 B1 来源于同一亲本染色体，它们之间是结合的；如果它们分别位于不同亲本的染色体上，则它们之间是排斥的。见图 13.6。

细胞周期相（phase of the cell cycle）：包括 G_1，S，G_2 和 G_0 期。见图 2.1。

内含子相（phase of an intron）：用于在编码序列中根据内含子中断信息的位置来划分内含子的名词。

表型（phenotype）：细胞或生物体可观察到的特征，包括除直接检测基因型外的任何检测结果。

系统发生（phylogeny）：根据所看到的进化上的亲缘关系来分类生物体。见图 12.22～12.24。

可塑性（plasticity）：见 transdifferentiation（转分化）。

多能性的（pluripotent）：严格来讲，指能发育为很多种细胞类型的能力，但与合子能发育为全部细胞类型不同。合子和其后立即形成的细胞被认为是全能的，但在囊胚阶段的分化意味着内细胞团的细胞是多能的而不是全能的。

点突变（point mutation）：引起某一基因座 DNA 序列微小变化的突变。含意有点含糊不清：当其与染色体突变相比较时，点突变的概念可能用来指单一基因内相当大的（是普通显微镜看不出来的）改变；而当涉及单一基因座的突变时，点突变通常是指仅一个核苷酸的替代、插入或缺失。

逐点 *p* 值（pointwise *p* value）：在连锁分析中，基因组中某一既定位置超过统计观察值的无义假设的概率。参照基因组范围 *p* 值。见节 15.3.4。

多腺苷酸化（polyadenylation）：附加典型的 200 个腺苷酸残基于 mRNA 3′端。Poly（A)尾对稳定 mRNA 十分重要。见图 1.18。

多基因的（polygenic）：由多个遗传基因座联合所决定的性状。数学上多基因的理论（节 4.4）假设有非常多的基因座，每个基因座都具有微效作用。

多态标记（polymorphic marker）：见框 8.1。

多态性（polymorphism）：严格意义上指群体中存在的 2 种或以上有意义频率的变异（等位基因、表型、序列变异、染色体结构变异）。分子遗传学家广义的使用包括（1）群体中频率＞1％的任何序列变异；（2）不考虑频率，任何非病理性的序列变异。

多倍体（polyploid）：遗传事件异常的（如组成的或嵌合的三倍体、四倍体等）或程序性（如一些植物和某些人体细胞是天然的多倍体）导致具有多个染色体组，即多倍体。

定位克隆（positional cloning）：在仅知道染色体位置条件下克隆基因。

引物（primer）：短的寡核苷酸，一般 15～25 个碱基，能与某一靶序列特异的碱基配对，从而使聚合酶启动互补链的合成。

原始生殖细胞（primordial germ cell）：胚胎或胎儿中存在的将最终发育为种系细胞的细胞。

探针（probe）：一个已知的 DNA 或 RNA 片段（或不同已知片段的集合），用于从一个复杂的、很不了解的核酸混合物中通过杂交分析鉴定密切相关的 DNA 或 RNA 序列。在标准杂交分析中探针被标记，反向杂交分析中靶序列被标记（框 6.4）。

启动子（promoter）：短序列元件的组合体，一般位于基因的上游，RNA 聚合酶与之结合启动基因的转录。见图 1.13。

校正读码（proofreading）：发现并校正 DNA 复制错误的酶学机制。

蛋白质截短实验（protein truncation test，PTT）：在一个联合的转录-翻译体系内通过人工表达一个突变等位基因来筛查链终止突变的方法。见图 17.9 和图 18.5。

蛋白质组（proteome）：一个细胞，组织或生物体内全部不同的蛋白质。参照基因组，转录物组。

原癌基因（proto-oncogene）：见癌基因。

原口动物（protostome）：见框 12.5。

染色体近端（proximal of chromosome）：位于相对靠近着丝粒的位置。

假常染色体区域（pseudoautosomal region）：位于 X 和 Y 染色体端，含有 X－Y 同源基因的区域。由于 X－Y 重组，这些区域内的等位基因表现出明显的常染色体遗传模式。见图 12.15。

假基因（pseudogene）：与非等位的功能基因序列高度同源但本身是没有功能的 DNA 序列。

焦磷酸测序（pyrosequencing）：一种适于检测与预先确定的起始点非常邻近序列的方法。见框 18.2。

数量性状基因座（quantitative trait locus，QTL）：在决定连续性状表型上很重要的基

因座。见节 15.6.8 讨论了潜在人类肥胖症主要 QTL 的探求。

cDNA 末端快速扩增法-聚合酶链反应：见框 5.1。

辐射杂种细胞（radiation hybrid）：在人类物理作图中，含有大量人类染色体小片段的啮齿类细胞。通过与致死剂量射线处理的人类细胞融合而产生。放射杂种嵌板可用作 STS 的快速定位。见图 10.4。

稀有切割工具酶（rare cutter）：由于其识别序列大和/或含有一个或多个 CpG 而偶尔切割 DNA 的限制性核酸酶，如 *Not* Ⅰ，*Sac* Ⅱ，*BssH* Ⅲ。见表 4.1。

读框（reading frame）：在翻译过程中，以一系列三联体密码子读取连续的 mRNA 序列的方式。任意的 mRNA 都有三种可能的读框，正确的读框是由 AUG 起始密码子的正确识别所决定的。

实时 PCR（real-time PCR）：见框 5.1。

隐性（recessive）：仅在纯合子才能表现的性状是隐性的。

重组体（连锁分析）［recombinant（linkage analysis）］：从双亲之一遗传了减数分裂交换形成的等位基因组合的人。

重组 DNA（recombinant DNA）：一个人工构建的包含不同来源共价连接序列的杂种 DNA，如带有插入片段的载体。

重组值（recombination fraction）：对于一对既定的基因座，发生因重组而分离的减数分裂的比例，通常以 θ 表示。θ 值介于 0 和 0.5 之间。见节 13.1。

重复 DNA（repetitive DNA）：基因组中以许多相同或相似拷贝存在的 DNA 序列。拷贝可以是串联的或分散的重复。

复制滑动（replication slippage）：串联重复 DNA 序列在复制过程中的错误，会导致新合成链与模板链相比额外多出或丢失重复单位。见图 11.5。

复制子（replicon）：能自我复制的任何核酸。许多克隆载体使用染色体外复制子（如质粒）。其他可直接使用染色体复制子（如酵母人工染色体），或者整合于染色体 DNA 而间接使用染色体复制子。

报道基因（reporter gene）：用于检验与其连接的上游序列引起其表达能力的基因。假定的顺式作用元件能被结合于一个报道基因，转染至适合的细胞来研究其功能。此外，在转基因动物（和其他生物体）制备中也常将无启动子的报道基因随机整合至染色体，使得报道基因的表达表明了一个有效启动子的存在。见框 20.4。

报道分子（reporter molecule）：与我们希望监测的 DNA 序列相连的，且易于检测的分子（例如荧光分子）。见图 6.7。

应答元件（response element）：通常位于启动子上游较近距离的，对细胞内环境的某些化学物应答而使基因表达的序列。表 10.4 列出了一些例子。

限制性片段长度多态性（restriction fragment length polymorphism，RFLP）：由于 DNA 序列多态性所致的含有不同大小的等位基因限制片段的遗传标记。见框 7.2。RFLP 最初由 DNA 印迹（图 7.5A）测定，现在一般通过 PCR（图 7.6）测定。

反转录转座子/反转座子（retrotransposon/retroposon）：借助于 RNA 中介物转座的一个转座 DNA 元件。反转录转座子编码一种作用于 RNA 转录物使其合成 cDNA 拷贝并整合至染色体 DNA 不同位置的反转录酶。见框 7.4，图 7.17 和图 7.18。

反转录病毒（retrovirus）：具有反转录酶活性的一种 RNA 病毒，在整合至宿主细胞染色体之前能使 RNA 基因组拷贝为 cDNA。

反转录酶（reverse transcriptase）：利用 RNA 模板合成 DNA 链的酶。用于构建 cDNA 文库（图 4.8）和 RT－PCR（节 20.2.4）。反转录是反转录病毒生命周期中必要的组成部分（图 18.2），但迄今已知不参与正常的细胞代谢过程。

RFLP：见限制性片段长度多态性。

核酶（ribozyme）：自然形成的或合成的催化 RNA 分子。见图 21.10。

RNA 编辑（RNA editing）：转录后 RNA 分子碱基序列发生特异性改变的一个自然过程，在人类基因很少发生，见图 8.16。

RNA 剪接（RNA splicing）：见剪接。

RNA 干扰（RNA interference，RNAi）：应用 siRNA 敲落（但很少是完全破坏）特异基因的表达。是研究基因功能强有力的工具。

罗伯逊融合（Robertsonian fusion）：将 2 条近端着丝粒染色体转变为一条中间着丝粒或亚中间着丝粒染色体的染色体重排。见图 2.21。有时也称为着丝粒融合，尽管实际上变化点位于短臂的近端而不是着丝粒。

反转录 PCR（reverse transcriptase PCR，RT-PCR）：见框 5.1。

基因表达的系列分析（serial analysis of gene expression，SAGE）：一种以测序为基础的表达谱分析方法。见图 19.5。

随体（染色体上）［satellite（on chromosome）］：可变地存在于端着丝粒染色体（13，14，15，21，22）短臂的柄状凸出物。

卫星 DNA（satellite DNA）：原意用来描述由于碱基组成不同而在密度梯度离心中形成独立的小带的 DNA 片段。该 DNA 由很长串的串联重复 DNA 序列组成。见节 9.4.1。

二级结构（secondary structure）：单链核酸或蛋白/多肽分子中距离较远的核酸或氨基酸间形成化学键所致的复杂结构区域。二级结构通常是由链内氢键形成所致（图 1.7 和图 1.24）。

节段非整倍体综合征［segmental aneuploidy（or aneusomy）syndrome］：见邻接基因综合征。

节段性复制（segmental duplication）：存在于不同染色体上或同一染色体内多个位置的高度相关的 DNA 序列板块。见图 12.12 和图 12.13。

分离分析（segregation analysis）：推断遗传模式的一种统计方法学。

分离比率（segregation ratio）：后代从亲代遗传某一既定基因或性状的比率。

半显性的（semi-dominant）：杂合子表型介于野生型和纯合子之间（但不一定是完全中间）的等位基因。此术语广泛用于小鼠遗传学，但至少在人类遗传学已很好地避免了，因为显性是一个性状特性而不是等位基因特性。

有义链（sense strand）：一个基因中与模板链（反义链）序列互补，并与转录的 RNA 序列一致（除了 DNA 含有 T 而 RNA 含有 U）的 DNA 链。引证的基因序列通常是有义链，按 5′→3′方向。见图 1.12。

序列同源性（sequence homology）：两个核酸序列或多肽序列间相似性的一种衡量

标准。

序列标签位点（sequence tagged site，STS）：设计的特异性 PCR 检测，可在任意 DNA 样本中容易地检测其有或无的任一独特的 DNA 片段。见框 10.3。

霰弹法测序（shotgun sequencing）：一个大克隆或整个基因组随机产生的 DNA 片段的测序。见图 8.3。

同胞对分析（sib-pair analysis）：见受累同胞对分析。

同胞（sibs）：兄弟或姐妹。

信号序列（前导序列）[signal sequence（leader sequence）]：位于一多肽 N 端、在细胞内或细胞外控制其目标的约 20 个氨基酸的序列。见节 1.5.4。

沉默子（silencer）：抑制基因转录组合的短 DNA 序列元件。见框 10.2。

沉默（同义）突变 [silent（synonymous）mutation]：改变密码子但不改变编码氨基酸的突变。见框 11.3。这样的突变在 mRNA 剪接或稳定性中仍起作用。

短散在核元件（short interspersed nuclear element，SINE）：一类中、高度重复 DNA 序列家族。在人类中最常见的是 Alu 重复家族。见表 9.15、图 9.17 和图 9.18。

小干扰 RNA（small interfering RNA，siRNA）：能特异地破坏与其同源的 mRNA 功能，长约 21～23nt 的双链 RNA。见节 20.2.7 和图 20.12。

简单重复序列多态性（simple sequence repeat polymorphism）：见微卫星。

姐妹染色单体交换（sister chromatid exchange，SCE）：涉及姐妹染色单体的重组事件。因为姐妹染色单体是彼此的复制品，如果交换不是不等的，则这样的交换不会有任何作用。但 SCE 频率的增加是 DNA 损伤的证据。

姐妹染色单体（sister chromatid）：存在于一条染色体内、以着丝粒相连接的两条染色体。非姐妹染色单体位于不同的但同源的染色体上。

位点专一诱变（site-directed mutagenesis）：DNA 序列中特定的预期改变的产物。可通过体外 DNA 克隆（节 5.5.2 和节 5.5.3）或体内同源重组来实现（节 20.2.6）。

滑链错配（slipped strand mispairing）：见复制滑动。

单核苷酸多态性（single nucleotide polymorphism，SNP）：单个核苷酸的任意多态性变异。SNP 包括 RFLP 以及其他不改变任何限制酶切位点的多态。虽然信息含量比微卫星多态少，但 SNP 更适合大规模、自动化分析。

体细胞杂种（somatic cell hybrid）：两种不同类型的体细胞，尤其是来自不同物种的细胞间融合形成的人工构建的细胞。人-啮齿类杂种细胞已成为有价值的定位工具。见框 8.4。

体细胞（somatic cell）：指除配子外机体内所有的细胞。

体节（somite）：节段中胚层中不连续的块。它通过形成大多数中轴骨（包括脊柱）、随意肌及部分真皮组织而建立身体的节段性组成。

DNA（Southern）印迹（Southern blot）：将 DNA 片段从电泳凝胶转移至尼龙膜或醋酸纤维膜（滤膜），准备进行杂交分析的方法。见图 5.12。

特异性（specificity）：衡量一种检测方法功效的标准。特异性＝（1－假阳性率），见图 5.12。

剪接受体位点（splice acceptor site）：内含子 3′端和下一外显子起始位置间的连接。一

致序列为 y11nyagR（y=嘧啶，R=嘌呤；大写字母=外显子）。见图 1.15。

剪接供体位点（splice donor site）：一个外显子末端和下游内含子 5′端起始位置间的连接。一致序列为（C/A）AG*gt*ragt（r=嘌呤；大写字母=外显子）。见图 1.15。

剪接体（spliceosome）：用于 RNA 剪接的核糖核蛋白复合体。

剪接（splicing）：一般指 RNA 剪接，即内含子转录的 RNA 序列从早期转录物上被切除掉，外显子转录的序列按照与外显子一样的线性顺序剪接到一起。一种 DNA 剪接形式对于 B 和 T 淋巴细胞中 T 细胞受体基因和多产的免疫珠蛋白的组装是重要的（图 1.14 和图 1.16）。

剪接增强子（splicing enhancer）：能提高邻近潜在剪接位点被实际使用概率的序列（可能是外显子的或内含子的）。见节 11.4.3 和节 16.4.1。

单链构象多态性（SSCP 或 SSCA）：单链构象多态性或分析。是点突变筛查常用的方法。见图 18.3A。

干细胞（stem cell）：可作为分化细胞的前体细胞，但仍保留自我更新能力的细胞。见节 3.4.3。

黏性末端 [sticky end（cohesive termini）]：双链 DNA 分子上短的单链突出物，常由某种限制性内切酶消化形成。具有互补黏性末端的分子在 DNA 连接酶作用下可共价结合形成重组 DNA 分子。见图 4.3 和图 4.4。

层化（stratification）：在假定同源的群体中存在的遗传上不同的组（群）。

STR 多态性（短串联重复序列多态性）（STR polymorphism，short tandem repeat polymorphism）：见微卫星。

序列标签位点（STS）：见框 8.1。

消减克隆（subtraction cloning）：鉴别在一个 DNA 样品中存在而在另一个大体相似的样品中缺乏的 DNA 序列的一种克隆方法。用于通过与广泛表达的 cDNA 文库消减选择组织特异性 cDNA，或者通过与正常 DNA 消减克隆在某一疾病患者中缺失的基因。

抑制型 tRNA（suppressor tRNA）：反密码子发生核酸替代突变的转运 RNA 分子。显示出改变的编码特异性并能翻译为一个无义（或错义）密码子。见框 5.3。

合胞体（syncytium）：没有细胞分裂的 DNA 复制循环或者多个细胞融合（如肌纤维细胞的情况）形成的含有多个细胞核的细胞。

同义（沉默）置换 [synonymous（silent）substitution]：一个密码子被另一个编码同一氨基酸的密码子所替代的置换。见框 9.2。

同线性（synteny）：基因座若位于同一条染色体上，则称为同线性的。同线性基因座不一定必须是连锁的。一条染色体上相距足够远的基因座随机分配，形成 50%重组体。

端粒（telomere）：染色体末端的一种特化的结构。人类的端粒由短串联重复序列阵（TTAGGG）*n* 组成，形成一个封闭的环，保护染色体末端。

模板链（template strand）：在转录中与新生成的 RNA 转录物碱基配对的 DNA 链。见图 1.12。

终止密码子（termination codon）：mRNA（延伸至一个基因）中标志多肽末端的 UAG（琥珀色）、UAA（黄褐色）或 UGA（乳色）密码子。

治疗性克隆（therapeutic cloning）：应用移植患者细胞核到人类 ES 细胞，治疗疾病的一种具有应用前景的方法。

反式作用（*trans*-acting）：一个调节因子不考虑染色体的位置影响靶基因所有拷贝的表达。反式作用调节因子通常是能识别靶位点的蛋白质。

转录单位（transcription unit）：以单一方式自然转录的、能产生单一初始转录物的 DNA 序列。真正的情况，一个转录单位就是同一个基因。

转录物组（transcriptomics）：一个细胞或组织中全部的不同 RNA 转录物。参照基因组、蛋白质组。

转分化（或可塑性）［transdifferentiation（or plasticity）］：似是决定一特定类型分化的细胞转入另一分化途径的可能性。

转导（transduction）：病毒介导的基因转移。见框 19.1 和 21 章伦理学框 1。

转染（transfection）：指真核细胞摄取外源 DNA（与细菌的转化意思相同，但在真核细胞中转化有不同的含义）。也指细菌细胞摄取质粒 DNA。见框 19.1。

转化（一个细胞的）［transformation（of a cell）］：1. 感受态细菌从环境中摄取裸露的高分子量 DNA；2. 正常真核细胞生长特性的改变，作为向肿瘤细胞演化的一个步骤。

转基因（transgene）：指被转入至动物或植物细胞中的外源基因。可存在于一些组织（如在人类基因治疗中）或所有组织（如在小鼠的种系工程中）。导入的转基因可能是游离的、瞬时表达的，或整合至宿主细胞染色体。

转基因动物（transgenic animal）：人工导入外源 DNA（转基因），并随后稳定整合于种系的动物。见图 20.2 和图 20.3。

转换（transition）：G↔A（嘌呤对嘌呤）或 C↔T（嘧啶对嘧啶）核苷酸替代。

易位（translocation）：非同源染色体间染色体区域的转移。见图 2.21。

传递不平衡检验（transmission disequilibrium test，TDT）：等位基因关联的一种统计学检验。见框 15.3。

转座子（transposon）：一种可动的遗传元件——见图 9.17。

颠换（transversion）：指嘌呤被嘧啶的核苷酸替代，反之亦然。

三倍体（triploid）：指具有三个基因组拷贝的细胞；或由三倍体细胞构成的生物体。

三体性复原（trisomy rescue）：有丝分裂不分离产生了成为整个胚胎祖先的二体性细胞，使最初的三体性胚胎存活。能导致单亲二体性。

三体性（trisomy）：某一条染色体具有三个拷贝。例如 21 三体。

滋养层［trophoblast（=trophectoderm）］：囊胚极性细胞的外层（图 3.13），将继续发育为胎盘的胚胎成分中的绒毛膜。

肿瘤抑制基因［tumor-suppressor gene（TSG）］：正常功能为抑制或控制细胞分裂的基因。TSG 在肿瘤中多失活。见节 17.4。

二次打击假说（two-hit hypothesis）：Knudson 的理论认为遗传性肿瘤需要单一细胞的两次连续突变。见图 17.5。

双杂交系统（two-hybrid system）：见酵母双杂交系统。

不等交换［unequal crossover（UEC）］：位于同源染色体的非姐妹染色单体上的非等位基因序列间的重组。见图 11.7。

不等姐妹染色单体交换（unequal sister chromatid exchange，UESCE）：一条染色体的姐妹染色单体上的非等位基因序列间的重组。见图 11.7。

单亲二倍体（uniparental disomy）：某一特定染色体对的两个拷贝来源于双亲之一的细胞或生物体。单亲二倍体可以是或不是病理性的，取决于所涉及的染色体。见节 2.5.4，框 16.6。

非翻译区（5′UTR，3′UTR）[untranslated region（5′UTR，3′UTR）]：mRNA 5′端在 AUG 翻译起始密码之前，或者 mRNA 3′端 UAG、UAA 或 UGA 终止密码子之后的区域。见图 1.19。

可变表达（variable expression）：具有某一既定基因型的人群中表型迹象的可变程度或强度。见例图 4.5C。

可变数目串联重复多态性 [variable number tandem repeat（VNTR）polymorphism]：微卫星、小卫星和卫星 DNA 是线性排列的串联重复序列，其重复单位的数目在人群中是可变化的。见框 7.2。VNTR 常指特定的小卫星。

载体（vector）：宿主细胞内能自我复制和自我维持的核酸序列，能使与其共价连接的任意序列具有相似的特性。

蛋白质印迹法（western blotting）：聚丙烯酰胺凝胶电泳分离不同大小的蛋白质，然后转移至硝酸纤维素膜，并用抗体进行检测的过程。见图 20.9。

全基因组扩增（whole genome amplification）：利用高度变性的引物扩增基因组中大量随机序列的 PCR 方法。可用于从单个细胞（如单个精子）中重复检测 DNA 序列（节 11.5.4）。

整装原位杂交（whole mount *in situ* hybridization）：见原位杂交。

X 失活（莱昂作用）[X inactivation（lyonization）]：由于专有的遗传印记形式所致的雌性哺乳动物细胞中两条 X 染色体中的一条失活。见节 10.5.6。

酵母人工染色体（Yeast artificial chromosome，YAC）：能够在酵母细胞中插入 Mb 或更大片段的一种载体。见图 5.17。

酵母人工染色体转基因的（YAC transgenic）：转入一完整的酵母人工染色体的转基因小鼠。用于研究被涉及的基因周围序列的调节作用。见节 20.2.3。

酵母双杂交系统（Yeast two-hybrid system）：用于鉴别、纯化与感兴趣的蛋白相结合的蛋白质的一个重要系统。见图 19.18。

锌指（zinc finger）：能稳定结合锌原子并使蛋白质具有与 DNA 序列特异性结合能力的多肽基序。常见于转录因子。见图 10.8。

动物印迹（zoo-blot）：含有来自不同物种的 DNA 样本的 DNA 印迹。见图 7.9。

合子（zygote）：受精的卵细胞。

（宫立国　译）

索　引